Addressing the Uncertainty Due to Random Measurement Errors in Quantitative Analysis of Microorganism and Discrete Particle Enumeration Data

by

Philip John Schmidt

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2010

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

ABSTRACT

The concentration of microorganisms (or discrete particles) in water is often evaluated by enumeration: the count obtained from a sample of specified volume is used to estimate the concentration. There are, however, several sources of random variability associated with the process of collecting enumeration data that can cause the count per unit volume to be a biased concentration estimate and that will make it an imprecise estimate. The actual concentration that is estimated using the available data is, therefore, uncertain. The error in concentration estimates is described as measurement error because the concentration cannot be measured exactly. Measurement error may include both unavoidable random errors in sample collection (*e.g.* randomness in the number of microorganisms contained in a sample) and analytical errors in counting the microorganisms in the sample (*e.g.* imperfect analytical recovery due to losses during sample processing or counting errors). To calibrate concentration estimates to actual microorganism concentrations, the count per unit volume must be divided by either the mean analytical recovery of the enumeration method or a sample-specific recovery estimate. Accordingly, appropriate analysis of enumeration data is dependent upon information about the analytical recovery of the enumeration method that is used.

The analytical recovery of an enumeration method is evaluated by seeding samples with known quantities of microorganisms or particles and then comparing the number that are observed in the enumerated sample to the number that were seeded. The fraction of seeded microorganisms that are observed, however, is an imprecise estimate of analytical recovery (particularly if the number of microorganisms seeded into the sample is not precisely known). It is demonstrated in this thesis that the standard deviation of such recovery estimates will be greater than the standard deviation of analytical recovery itself because of measurement error in the recovery estimates. Accordingly, the effect of the seed dose (and the precision thereof) upon the precision of the recovery estimates must be addressed in experiments that are used to quantify analytical recovery (and the variability therein). Additionally, approaches that are used to analyze recovery data must appropriately address the measurement error associated with recovery estimates. Probabilistic models are developed herein to describe the variability in recovery estimates as a function of the seed dose and the variability in analytical recovery itself. These models are used to facilitate analysis of alternative recovery experiment designs so that experiments can be designed to yield adequately precise estimates of analytical recovery (or the mean and standard deviation thereof). Additionally, the probabilistic models are used to develop statistical tools that enable analysis of recovery data with appropriate regard for measurement errors. Direct use of conventional hypothesis tests and confidence intervals to analyze recovery estimates is often inappropriate because recovery estimates are often not normally distributed and may have non-constant error. Moreover, it is demonstrated in this thesis that such statistical tools will yield biased estimates of the standard deviation of analytical recovery (due to the effects of measurement error) and this will reduce the power of hypothesis tests to classify an obtained difference between the mean (or the difference between two means) and the null hypothesis as statistically significant. It is imperative to use statistical tools that enable appropriate analysis of the available recovery data because proper analysis of microorganism concentration data depends upon appropriate quantification of analytical recovery.

In this thesis, a statistical framework (using probabilistic modelling and Bayes' theorem) is developed to enable appropriate analysis of microorganism concentration estimates given information about analytical recovery and knowledge of how various random errors in the enumeration process affect count data. This framework is used to address several problems: (1) estimation of a single concentration value and quantification of the uncertainty therein from single or replicate data (possibly including non-detect samples), (2) estimation of the log-reduction of a treatment process (and the uncertainty therein) that is estimated by comparing pre- and post-treatment concentrations, (3) quantification of random concentration variability over time from temporally distributed enumeration data, and (4) estimation of the sensitivity (*i.e.* probability that microorganisms will be detected) of enumeration processes given knowledge about the associated measurement errors and analytical recovery. Each of these problems is of interest in drinking water treatment and research, and in Quantitative Microbial Risk Assessment (QMRA).

Investigation of the contemporary strategies that are used to analyze temporally variable pathogen concentrations in Monte Carlo QMRA has revealed that measurement errors in concentration estimates and the analytical recovery of the enumeration method (if addressed at all) are often addressed improperly and in ways that will result in bias (*e.g.* over-estimated risks). In contrast, the Bayesian framework that is developed within this thesis is a robust and appropriate strategy to address variability in pathogen concentrations (and the measurement errors therein) in Monte Carlo QMRA. Estimation of the sensitivity of an enumeration-based detection method is useful in the context of water treatment, but it is also particularly important in the analysis of errors in medical and epidemiological diagnoses. A statistical approach is developed herein that uses information about the analytical recovery of the enumeration method (and not just the relative frequency of non-detects) to rigorously analyze sensitivity.

Probabilistic models that describe the sources of random error in the enumeration process are not only useful to develop appropriate quantitative analysis approaches; they can also be used to evaluate the design of experiments. Simple probabilistic models and variance decomposition are used herein to develop experimental design guidelines for recovery experiments and for collecting more reliable microorganism concentration estimates. In the latter case, it is demonstrated that sample volumes should be chosen such that samples will typically contain at least 10 microorganisms in order to obtain acceptably reliable concentration estimates. It is also demonstrated that improving the analytical recovery of enumeration methods (*e.g.* reducing losses or the variability in analytical recovery) does not always have an appreciable effect upon the precision of associated concentration estimates. Therefore, method development should focus on providing inexpensive, efficient, and convenient methods that enable enumeration of large sample volumes rather than upon small improvements in analytical recovery.

This research demonstrates that probabilistic modelling that addresses random measurement errors in the enumeration process is a powerful tool to facilitate appropriate quantitative analyses in many different applications that are important in the water treatment industry. It also enables evaluation of the design of experiments so that more informative data can be obtained using the available resources.

ACKNOWLEDGEMENTS

In completing this research, I have been blessed with knowledgeable and helpful people, fun and educational opportunities, and excellent resources. I cannot fully explain my gratitude for all the things that have enabled me to do this work, but I will try!

My supervisors, Dr. Monica Emelko and Dr. Mary Thompson, have been incredibly helpful in supporting my work and inspiring me to be a rigorous researcher. Monica identified the need for this research in the water treatment industry (and beyond) and gave me the resources to pursue it. She is the first to congratulate me on a job well done, yet she always reminds me that there are things I can do better and so many more ideas to pursue. She also reminds me that there is so much more to a full academic life than research: we must teach the minds of tomorrow and volunteer our time and abilities to give back just a little bit of what we have received. We must also refresh our minds through time with friends and family and enjoying the great outdoors. Mary has been a great source of encouragement, has helped me to be mathematically correct, and has contributed many ideas to this research. She always smiles enthusiastically and gives encouragement when it is needed most. I cannot thank both supervisors enough!

This research could not have happened without scholarships provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), additional funding from the Canadian Water Network (CWN), and support in the form of scholarships, teaching assistantships, and research assistantships from the University of Waterloo.

I have been privileged to have the opportunity to work with Dr. Park Reilly. He has been an excellent mentor and teacher and is always available to discuss subtle details of Bayesian statistics. Thanks are also extended to Dr. Maria Mesquita for her microbiological knowledge and friendship. I am grateful for the contributions of all members of my examining committee. In particular, I thank my external examiner Dr. Charles Haas. It means a lot to have such an expert in the fields of environmental engineering, statistics, and microbiology take the time to review my research. I also acknowledge the efforts of the anonymous reviewers of my various publications: they have helped me to be more rigorous in many ways and to provide a didactic presentation of my research. Thanks to all the professors, administrators, and laboratory technicians who have enriched and enabled my experience at the University of Waterloo. I am grateful for the many opportunities to teach excellent students; these have enhanced my understanding of theory and my ability to communicate it. I am thankful for opportunities provided by Dr. Emelko to present my research in the far corners of North America and in beautiful Porto Alegre, Brazil, and to have met so many great researchers. Obrigado!

I am thankful for so many grad school friends who have enriched my life personally and professionally. In particular, I thank Dr. Trish Stadnyk, Lawrence Lee, Dr. Peter Knappett, Heather Sim, Bo Cheyne, and all the past and present students in Dr. Emelko's research group. I also thank Craig and Jane Campbell, Anne MacDonald, and many others for their friendship and support. I am grateful for the love and support of my parents, Ron and Elaine Schmidt, and other family members. My fiancée, Denise Mortley, and her family have also been incredibly supportive. Denise contributed greatly to the completion of this thesis by refusing to marry a student!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

This list of symbols summarizes the symbols that are used in the models throughout the main body of the thesis. Symbols that are used only briefly (*e.g.* in a single section or equation) are excluded from this list. More detailed definitions for specific models are included in Appendix B.

a	shape parameter of the beta distribution describing non-constant analytical recovery
b	shape parameter of the beta distribution describing non-constant analytical recovery
c	the concentration of analyte particles in the source
c^*	a generated concentration value from the posterior predictive concentration distribution
i	the sample index for replicate recovery or concentration data
	the sampling event index for temporally distributed concentration data
k	the sample index for replicate samples in a log-reduction analysis
	the sample index within a sampling event for temporally distributed concentration data
n	the precisely known number of seeded particles in a recovery sample
	the number of indigenous particles collected in a sample
n^*	the precisely known number of internal seed analyte particles
p	the probability (or rate) of recovery
p^*	a generated recovery value from the posterior predictive recovery distribution
r	the number of replicate samples
r_i	the number of samples in the i^{th} sampling event for temporally distributed concentration data
V	the sample volume
x	the number of analyte particles observed in a sample
x^*	the number of internal seed analyte particles observed in a sample
α	shape parameter of the gamma distribution describing non-constant analytical recovery
β	scale parameter of the gamma distribution describing non-constant analytical recovery
λ	the expected number of seeded particles (stock concentration × volume) in a recovery sample
	scale parameter of the gamma distribution describing temporal concentration variability
μ_c	mean concentration
μ_p	mean analytical recovery
θ	the fraction of sample that is enumerated in partial sample analysis
ρ	the log-reduction of a treatment process
	shape parameter of the gamma distribution describing temporal concentration variability
σ_p	standard deviation of concentration
σ_p	standard deviation of analytical recovery
Γ	the gamma function $\Gamma(.)$

Chapter 1
Introduction

Individual microorganisms and some types of particles can be regarded as discrete because they only occur in whole number quantities (*i.e.* it is not possible, or practically relevant, to have a fraction of a discrete object). Analytical methods that quantify specific types of analyte particles (*e.g.* microorganisms or discrete particles) by counting are classified as enumeration methods. Enumeration methods often involve considerable sample preparation to enable counting of the analyte particles, and the counting process itself is sometimes complicated (particularly if the analyte particles are obscured by other matter or if they are not easily identifiable). It is common for enumeration methods to yield counts that are different from the respective numbers of analyte particles that are actually present, and such methods can be said to have imperfect analytical recovery. Moreover, when enumeration is used to quantify the density of analyte particles in (or on) a particular medium, the density within the sample is unlikely to be exactly equal to the density in the source. This error will occur because analyte particles are discrete and the number contained in a sample of particular size is random even if the density in the source (*e.g.* the concentration in water) is homogeneous. This error causes the count of analyte particles per unit volume to be an imprecise concentration estimate even if the enumeration method enables the analyte particles to be counted precisely. Analytical recovery, “the capacity of the analyst to successfully count each microorganism or particle of interest in a sample using a specific enumeration method” (Schmidt *et al.*, 2010a), is also important because failure to count the analyte particles in samples precisely can lead to inaccurate (*i.e.* biased) concentration estimates and because variability in analytical recovery among samples adds to the variability of concentration estimates. The errors that make concentration estimates imprecise or inaccurate are described herein as measurement errors because the concentration of microorganisms (or discrete particles) in water cannot be measured exactly.

The concentration of a particular type of microorganism or discrete particle in water is typically estimated as the count obtained divided by the volume analyzed, but unavoidable measurement errors make these concentration estimates imprecise and potentially inaccurate. Inaccuracy can be resolved by dividing the count per unit volume by analytical recovery, but the sample-specific analytical recovery of the enumeration method is unknown and such a correction does not resolve the imprecision of the concentration estimate. It is important to recognize that all concentration estimates are uncertain and that proper assessment of this uncertainty must consider what the concentration might actually be given the available enumeration data and information about the measurement errors therein. For example, it may be found that a particular count of microorganisms or particles in a particular sample volume could have resulted from a wide range of concentration values (because of measurement errors). A thorough quantitative analysis of enumeration data must address uncertainty in concentration estimates (due to measurement errors) because assuming that concentration estimates are unbiased and precise when they are not can lead to incorrect interpretation of the data. A count of zero, for example, does not conclusively indicate that the analyte particles are absent in the source because unavoidable measurement errors can cause random non-detect results when analyte particles are actually present. It is important in many applications to properly consider what the concentration might actually be when a non-detect result is obtained. Measurement error must also be considered when evaluating the variability in concentration among a set of temporally distributed enumeration data. This variability can be obscured by measurement errors because the imprecision of unbiased concentration estimates will cause the estimates to be more variable than concentration itself. Therefore, it is important to consider how to analyze enumeration data in ways that appropriately address measurement errors, and how the results of such analyses would compare to analyses that ignore measurement errors. Additionally, it is helpful to consider whether or not possible changes in the design of future experiments will appreciably reduce measurement errors (*i.e.* enable collection of more informative enumeration data).

1.1 RESEARCH CONTEXT

Microorganisms are often quantified in water to determine if the water is microbiologically safe for its intended use by humans. The intent of such analyses is to reduce human exposure to pathogenic microorganisms that may cause infection, illness, or even death. Pathways for human exposure to waterborne pathogens include consumption of contaminated drinking water, contact with contaminated recreational water, and consumption of produce that was irrigated or washed with contaminated water. Accordingly, microorganisms that are potentially pathogenic or that indicate the potential presence of pathogens are often quantified in water, and the efficiency with which treatment processes reduce concentrations of various microorganisms is of interest. In the context of drinking water treatment, microorganisms may be enumerated to evaluate their occurrence in the source water (*e.g.* LeChevallier *et al.*, 1991a; Rose *et al.*, 1991b; Haas and Rose, 1996), in treated drinking water (*e.g.* LeChevallier *et al.*, 1991b; Rose *et al.*, 1991b), or in distribution systems (*e.g.* Christian and Pipes, 1983), or to investigate their occurrence and transport in the environment (*e.g.* Hansen and Ongerth, 1991; Mons *et al.*, 2009). Enumeration of microorganisms may be required by drinking water regulations (*e.g.* USEPA, 2006), or conducted to facilitate a risk assessment (*e.g.* Teunis *et al.*, 1997; Barbeau *et al.*, 2000; Masago *et al.*, 2002; Medema *et al.*, 2003; Petterson *et al.*, 2007; Jaidi *et al.*, 2009; Cummins *et al.*, 2010). Microorganisms (or discrete particles that are used as surrogates for microorganism transport) may also be enumerated before and after a particular bench-, pilot-, or full-scale treatment process (*e.g.* Rice *et al.*, 1996; Dugan *et al.*, 2001; Huck *et al.*, 2001; Emelko and Huck, 2004; Brown and Cornwell, 2007; Assavasilavasukul *et al.*, 2008) to evaluate treatment efficiency. *Cryptosporidium* oocysts have also been enumerated in the guts of neo-natal mice to evaluate the infectivity of oocysts in water samples following various types of disinfection (*e.g.* Peeters *et al.*, 1989; Korich *et al.*, 1990).

There are many other applications in which microorganisms are enumerated: medical applications include the enumeration of specific types of cells in blood (*e.g.* Sartor *et al.*, 2005) and epidemiological applications include the enumeration of parasites in feces (*e.g.* Teixeira *et al.*, 2007). Other applications include recreational water testing (*e.g.* Kinzelman *et al.*, 2003; USEPA, 2004) and evaluation of methods used to detect biological weapons (*e.g.* Edmonds *et al.*, 2009). In addition to various types of water samples, *Cryptosporidium* oocysts have been enumerated in feces (*e.g.* Kuczynska and Shelton, 1999; Pereira *et al.*, 1999; Robinson *et al.*, 2008), in soil (*e.g.* Kuczynska and Shelton, 1999; Davies *et al.*, 2005), in sewage sludge (*e.g.* Iacovski *et al.*, 2004), on produce such as lettuce and raspberries (*e.g.* Cook *et al.*, 2006), and in shellfish (*e.g.* Downey and Graczyk, 2007).

It is not possible to comprehensively address the measurement errors and data analysis strategies associated with all the different ways (and applications) in which enumeration data are generated. This thesis focuses mainly upon enumeration of microorganisms or discrete particles in water samples associated with the production of microbiologically safe drinking water. The research is not specific to any one type of method, microorganism, or particle. It applies to any method in which each water sample is processed to yield a single count. Multiple counts from a single sample (*e.g.* dilution series, split samples) would not be statistically independent and would require more complicated and case-specific analysis approaches, while multiple independent samples from the same source are addressed as replicates. Much of the discussion in this thesis focuses upon the enumeration of *Cryptosporidium* oocysts in water samples because (1) it is an important waterborne pathogen (*e.g.* Craun *et al.*, 1998), (2) enumeration methods and quality control criteria have been strictly standardized (*e.g.* USEPA, 2005a; USEPA, 2005b), (3) it is specifically addressed in important water treatment regulations (*e.g.* USEPA, 2006), and (4) there is considerable existing literature addressing evaluation of its analytical recovery or concentration in various types of water. Outside the immediate focus of this thesis, many of the concepts and conclusions that are discussed are generally applicable to many types of microorganisms and discrete particles, to many types of enumeration methods, and to many types of media from which samples are collected and enumerated. Accordingly, the concepts presented herein should be considered in many applications in which enumeration data are obtained and possibly also in other methods that are used to detect and quantify specific types of microorganisms (*e.g.* most probable number methods, polymerase chain reaction methods), but detailed discussion of these implications is beyond the scope of this thesis.

1.2 RESEARCH APPROACH

Statistical analysis is a framework to draw scientifically valid conclusions and to facilitate sound management decisions despite the uncertainty that arises from data that are limited in quantity and precision. Proper statistical methods for data analysis arise from an appropriate understanding of how the data (*e.g.* counts) are generated with respect to the parameters (*e.g.* concentration, analytical recovery) that are of interest; statistical methods that fail to address this can yield misleading results by violating important assumptions. In such cases, it is unclear whether the data analysis facilitates informed decisions in light of the available data, or whether it is merely an exercise in manipulating the data to obtain a result that is inconsistent with the asserted meaning: “It has been alleged that certain people use statistics as a drunk does a lamppost – more for support than illumination" (Eisenhart and Wilson, 1943).

The central theme of this thesis is the development of statistical methods to analyze microorganism and discrete particle enumeration data (particularly in drinking water applications) with appropriate consideration of the sources of variability (including random measurement errors) that affect enumeration data. This is completed through the development of probabilistic models (or selection of existing probabilistic models where appropriate) that describe how the enumeration data are generated (in a mathematical sense) as a function of the parameter(s) of interest and various sources of variability. Such modelling must be done with careful consideration of all assumptions (both practical and mathematical). Accordingly, the assumptions are clearly explained (and evaluated where possible), and the consequences of violating them are described. Once a probabilistic model has been proposed to model the effects of various sources of variability upon enumeration data in a particular application, it can be used in two general ways: to investigate the distribution of the data (or some function of the data) as a function of known parameters, or to evaluate the uncertainty in unknown parameters given available data. The former type of analysis can be used to aid design of experiments that will enable collection of more precise recovery and concentration estimates, and is addressed herein using Monte Carlo simulation and variance decomposition. The latter type of analysis enables a quantitative description of what the analyst is entitled to believe about a parameter (*e.g.* analytical recovery or variability thereof, concentration or variability therein, log-reduction of a treatment process, or sensitivity of an enumeration method) given data and knowledge about measurement errors. This is addressed herein using maximum likelihood estimation and Bayes' theorem, and the results of such analyses are compared to various alternative approaches that do not address measurement errors.

1.3 RESEARCH OBJECTIVES

Parameters associated with the enumeration of microorganisms and discrete particles in water, such as the analytical recovery of an enumeration method, the microorganism or particle concentration in a source, the log-reduction of a treatment process, and the sensitivity of an enumeration method, cannot be measured exactly, yet scientific conclusions and drinking water management decisions are often made without considering the uncertainty in the *measurements* upon which they are based. The principal objective of this research is to facilitate analysis of enumeration data that addresses random measurement errors, which will enhance the statistical rigor of scientific analyses and enable the drinking water industry to make decisions with appropriate regard for unavoidable measurement errors. The specific objectives that are addressed in pursuit of this general goal are listed below and are addressed in Chapters 4-8 respectively.

1) To quantify the variability in the analytical recovery of an enumeration method with appropriate consideration of the random measurement error in recovery data.

2) To develop a robust approach to quantify the uncertainty (due to random measurement errors) that is associated with estimates of the concentration of microorganisms and discrete particles in water and the log-reduction achieved using a particular water treatment process.

3) To incorporate information about the uncertainty due to random measurement error in pathogen concentration estimates into Monte Carlo Quantitative Microbial Risk Assessment.

4) To develop rigorous protocols for interpreting the information about concentration that can be obtained from non-detect samples, and for estimating the sensitivity of an enumeration method as a function of random measurement errors and analytical recovery data.

5) To identify experimental design strategies that may enable collection of more precise recovery and concentration estimates by reducing the most impactful random measurement errors.

1.4 THESIS ORGANIZATION

Chapter 2 provides a summary of random errors in the enumeration of microorganisms or discrete particles in water and defines the associated terminology (including analytical recovery) that is used throughout this thesis. Chapter 3 introduces the statistical concepts and tools that are used (or addressed) in this thesis, with emphasis upon probabilistic modelling and Bayes' theorem. Chapter 4 addresses strategies that are used to quantify the analytical recovery of an enumeration method, probabilistic modelling of random errors for various types of recovery experiments, experimental design considerations associated with recovery experiments that have replicate samples, and various statistical approaches to interpret recovery data. Chapter 5 addresses the evaluation of microorganism or discrete particle concentrations in water. In addition, it discusses the dispersion of analyte particles in water (*e.g.* whether they are clumped or randomly distributed) and the comparison of two concentrations to estimate the log-reduction of a treatment process. The chapter includes a review of literature that has addressed random errors in enumeration-based concentration and log-reduction estimates using probabilistic models. Chapter 6 focuses upon modelling and quantitative analysis of situations in which enumeration data are obtained on numerous occasions at a specific location and in which the temporal concentration variability is of interest. The situation considered in this chapter is related to regulatory monitoring of *Cryptosporidium* oocysts in the United States' Long Term 2 Enhanced Surface Water Treatment Rule (USEPA, 2006). It is also related to Monte Carlo Quantitative Microbial Risk Assessments (QMRA) in which the temporal variability of pathogen concentrations in the water is of interest. A review of QMRA literature pertaining to *Cryptosporidium* oocysts or *Giardia* cysts in water is provided, and the modelling approaches that have been used (particularly concerning analytical recovery) are evaluated. Chapter 7 addresses the analysis of non-detect samples (particularly with respect to the method detection limit) and provides a case study analysis in which the sensitivity of an enumeration method is estimated from analytical recovery data. Chapter 8 uses mathematical expectation and variance decomposition to evaluate the mean and variance of enumeration data (or functions thereof). The results are used to mathematically prove various types of bias and to evaluate various strategies that may enable the collection of more informative enumeration data. Chapters 4 through 8 are each concluded with a discussion of the practical implications resulting from the presented research. Chapter 9 provides a summary of the conclusions and implications that arise from the research presented in this thesis as well as recommendations for future research.

Chapter 2

Sources of Variability in the Collection and Enumeration of Water Samples

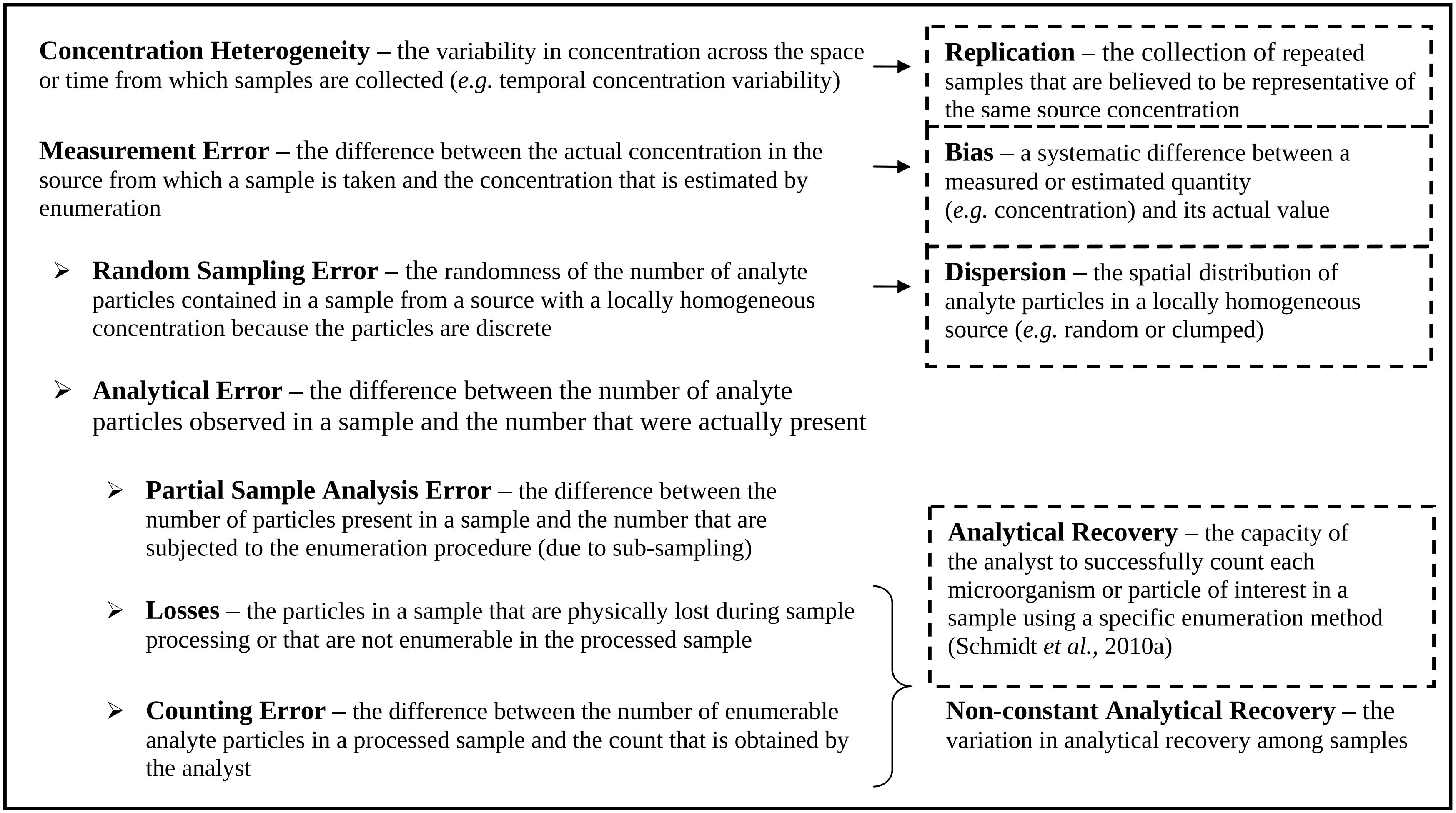
The concentration of a specific type of analyte particles (*i.e.* microorganisms or discrete particles) in water cannot be measured directly, but is often of interest. The count of analyte particles in a representative sample drawn from the source is often used to estimate the concentration. It has generally been found that such counts are highly variable, even if repeated samples are drawn from the same homogeneous source. The variability among a set of enumeration data can be attributed to several sources of variability ('errors') in the process of collecting and enumerating samples. These errors can occur in a well controlled process and are not to be confused with mistakes that lead to non-representative data. This chapter is comprised of three sections: a summary of the types of errors that can lead to variable enumeration data when evaluating the analyte particle concentration in water (Section 2.1), discussion of possible errors in the enumeration of a sample (Section 2.2), and definition of the term analytical recovery with respect to enumeration methods (Section 2.3).

2.1 ERRORS IN ENUMERATION DATA

Counts of microorganisms or discrete particles can vary because the actual concentration in the source varies spatially or temporally among the samples or because there are errors in the sample collection and enumeration process. The former is defined herein as concentration heterogeneity while the latter is defined as measurement error (because the concentration at a particular location and time is not measured exactly). Measurement error can result from variability associated with sample collection (random sampling error) or from errors in counting the analyte particles in a sample (analytical errors). Analytical error can occur because of errors in sample processing that change the number of enumerable analyte particles (*e.g.* losses or analysis of only a portion of the sample), or due to errors in counting the analyte particles in the prepared sample. Analytical error (except for the error due to partial sample analysis) is related to the analytical recovery of the enumeration method. A final source of error is that the analytical recovery of the enumeration method may vary among samples (an error that is referred to as non-constant analytical recovery). Figure 2.1 summarizes these errors along with several relevant terms.

2.1.1 Concentration Heterogeneity

Concentration heterogeneity concerns the spatial or temporal variability in concentration throughout a source from which samples are drawn. Unless repeated samples are collected very close in space and



Concentration Heterogeneity – the variability in concentration across the space or time from which samples are collected (*e.g.* temporal concentration variability)

Replication – the collection of repeated samples that are believed to be representative of the same source concentration

Measurement Error – the difference between the actual concentration in the source from which a sample is taken and the concentration that is estimated by enumeration

Bias – a systematic difference between a measured or estimated quantity (*e.g.* concentration) and its actual value

- **Random Sampling Error** – the randomness of the number of analyte particles contained in a sample from a source with a locally homogeneous concentration because the particles are discrete

Dispersion – the spatial distribution of analyte particles in a locally homogeneous source (*e.g.* random or clumped)

- **Analytical Error** – the difference between the number of analyte particles observed in a sample and the number that were actually present
    - **Partial Sample Analysis Error** – the difference between the number of particles present in a sample and the number that are subjected to the enumeration procedure (due to sub-sampling)
    - **Losses** – the particles in a sample that are physically lost during sample processing or that are not enumerable in the processed sample
    - **Counting Error** – the difference between the number of enumerable analyte particles in a processed sample and the count that is obtained by the analyst

Analytical Recovery – the capacity of the analyst to successfully count each microorganism or particle of interest in a sample using a specific enumeration method (Schmidt *et al.*, 2010a)

Non-constant Analytical Recovery – the variation in analytical recovery among samples

Figure 2.1: Types of errors and important considerations in the collection of enumeration data

The hierarchical structure of errors in the collection of enumeration data is shown on the left (*i.e.* concentration estimates can vary due to concentration heterogeneity or measurement errors, measurement errors consist of random sampling error and analytical error, and analytical error can include partial sample analysis error, losses, and/or counting error). Important considerations associated with these errors are shown on the right in dashed boxes. Non-constant analytical recovery is an additional error representing variation in the degree of losses and counting errors among samples.

time (or the source is well-mixed and at steady state), the concentration of analyte particles at the location and time from which each sample was drawn may vary among samples (*i.e.* it may not be appropriate to regard the samples as replicates). In this thesis, it is assumed that concentration heterogeneity is absent when concentration estimates are regarded as replicates (Chapter 5) and random temporal concentration variability is considered in Chapter 6. Non-random concentration heterogeneity can occur if the concentration varies among samples according to a pattern (*e.g.* if there are spatial or temporal trends or autocorrelation) so that the associated concentrations are not independent and identically distributed. Complex spatial or temporal regression models are not addressed herein, and it is assumed throughout that samples collected from heterogeneous sources are appropriately representative of the concentration being measured (*i.e.* representative of the bulk average across the space and/or time being described by the resulting concentration estimate). The term 'representative sampling error' was used in Emelko *et al.* (2010a) to describe concentration heterogeneity because the presented models assumed that each sample was representative of the concentration that the analyst wished to quantify and that repeated samples were replicates.

2.1.2 Random Sampling Error

Random sampling error ('Student', 1907; Nahrstedt and Gimbel, 1996) addresses randomness of the number of analyte particles contained in a sample from a source with a locally homogeneous concentration because the particles are discrete. It has also been described as "the variability in the number of particles contained in replicate samples of equal size that are drawn from a homogeneous source" (Emelko *et al.*, 2010a). It is not possible for discrete particles to be blended evenly throughout the source and the quantity of particles in a sample is necessarily a whole number. Accordingly, the number of particles in a sample depends on which portion of the source is sampled and how many particles are randomly present there. Random sampling error depends upon the dispersion of analyte particles in the water (*i.e.* whether they are randomly scattered, clustered, or somewhat evenly spaced). Further discussion of dispersion is included in Section 5.2.

2.1.3 Analytical Error

Analytical error describes the discrepancy between the quantity of analyte particles contained in a sample and the resulting number of observed particles because of imperfect analytical recovery (Nahrstedt and Gimbel, 1996; Emelko *et al.*, 2010a) or partial sample analysis. These errors associated with the processing and enumeration of a sample are discussed in greater detail in Section 2.2. Analytical recovery corresponds to losses during sample processing and errors in counting the processed sample. Its definition is discussed in Section 2.3, and Chapter 4 addresses quantification of the analytical recovery of an enumeration method. Partial sample analysis addresses the situation in which only a portion of a sample is enumerated. Most of the models presented in this thesis assume that the entirety of each sample is enumerated, but partial sample analysis is addressed in Sections 4.2.5, 5.1.3, 8.1.3, and 8.2.1. In general, if the fraction of the sample that was enumerated is known and the sample was well-mixed when it was subdivided, then the error due to partial sample analysis can be incorporated into concentration analyses as a reduced sample volume (*i.e.* the ‘effective’ or ‘equivalent’ volume) (Parkhurst and Stern, 1998; Young and Komisar, 1999).

2.1.4 Non-constant Analytical Recovery

Analytical recovery may vary among samples due to controllable or measurable factors or due to apparent randomness (*i.e.* uncontrollable or immeasurable factors). A distribution is used to describe non-constant analytical recovery (*e.g.* Nahrstedt and Gimbel, 1996) when it is assumed that the method is in a state of statistical control (Schmidt *et al.*, 2010a). Statistical control is discussed in Section 3.1.1.

2.1.5 Counting Error

Counting error describes the discrepancy between the quantity of enumerable analyte particles in a processed sample and the resulting number of observed particles. This error may arise from under- or over-counting the enumerable analyte particles in the processed sample or from identification errors (*i.e.* falsely counting other types of particles or falsely regarding an analyte particle as something else) (Emelko *et al.*, 2010a). Counting errors may depend on when, or by whom, the processed sample is enumerated. Throughout this thesis, counting errors are regarded as a component of analytical error that is implicitly included in recovery data. Non-random counting errors (*i.e.* substantial analyst mistakes) are assumed to be minimized by suitable quality control and analyst capability. Counting errors are described in greater detail in Section 2.2.

2.2 ANALYTICAL ERRORS IN ENUMERATION METHODS

Analytical error in an enumeration method represents the difference between the number of analyte particles observed in a sample (or in a processed portion of sample) and the number that were actually present in the sample. The difference between the actual and observed quantities of analyte particles is due to imperfect analytical recovery (*e.g.* unintentional loss of particles or error in enumerating the processed sample) or partial sample analysis (*i.e.* enumeration of only a portion of the sample). In assessing analytical errors, it is helpful to consider the enumeration process as a system in which every analyte particle has a source and a sink (*e.g.* Figure 2.2). In this figure, partial sample analysis is represented by sub-sampling because some number of the particles are not subjected to the entire enumeration process. Errors associated with sample enumeration are considered separately from other errors in this system, but are also regarded as analytical errors. Collectively, analytical errors necessitate calibration of observed quantities of analyte particles to the actual quantities that are of interest. The various types of errors that are depicted in Figure 2.2 are discussed below. This section is concluded with a case study of two *Cryptosporidium* oocyst enumeration methods and the possible errors associated with sample processing and enumeration (Section 2.2.5).

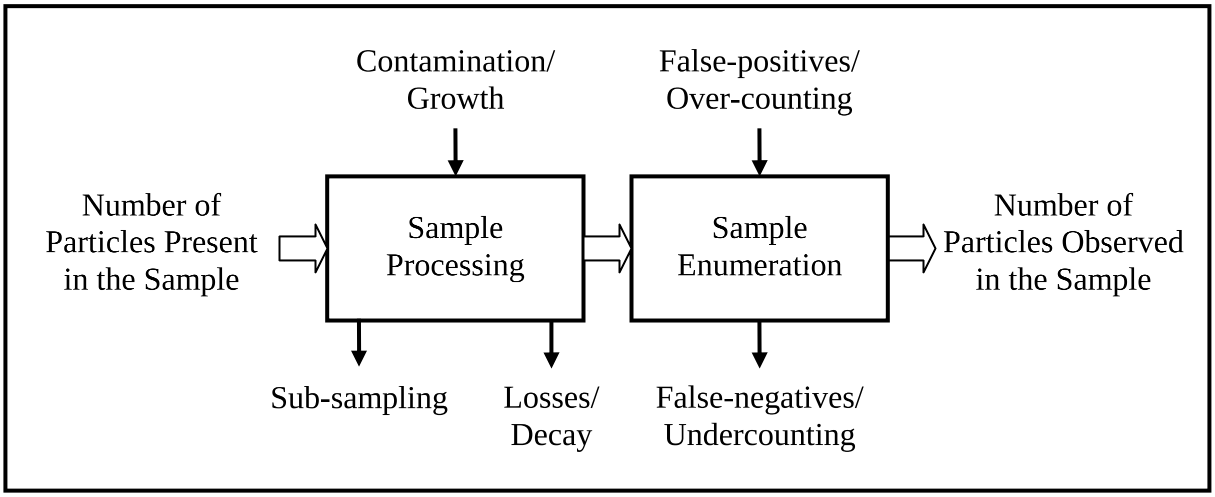


Figure 2.2: Analytical errors in enumeration methods

2.2.1 Increases in the Number of Analyte Particles

The quantity of analyte particles in a processed portion of sample can only be increased by contamination (*i.e.* introduction of particles that are not indigenous to the sample due to improperly cleaned apparatus or cross-contamination of samples) or, in the case of microorganisms, by growth. It is assumed in this thesis that contamination is a non-random error (*i.e.* an error that cannot be modelled using a distribution) that can be suitably minimized through appropriate procedures for processing samples and cleaning equipment. Failure to control contamination compromises the value of enumeration data because obtained counts may be more closely related to the unquantified contamination than to the actual quantity of analyte particles in the sample. Likewise, growth in the number of enumerable microorganisms during sample processing must be minimized by appropriate sample processing procedures because it is often a non-random error (*i.e.* it is likely not possible to predict, according to any consistent model or distribution, how much growth there may have been). In culture-based methods, growth is used as a means of detecting microorganisms (*e.g.* colony- or plaque-forming units), but it must be assumed that the quantity of culturable microorganisms did not increase between sample collection and incubation. The models presented in this thesis assume that the number of analyte particles cannot be increased: particles can only be lost or counted incorrectly.

2.2.2 Losses of Analyte Particles

There are often many opportunities during sample processing for analyte particles to become lost. In aqueous samples, there can be losses any time the fluid is transferred from one vessel to another. These losses can include attachment to apparatus (*e.g.* tubing, filter manifolds, pipettes), losses in the sample vessel (*e.g.* residual sample in, or particles attached to, the vessel), and spills or leaks. There can also be losses during sample concentration and purification (*e.g.* particles discarded in the filtrate or supernatant, or not successfully eluted from a filter). Sample processing may also cause particles to disintegrate or cause microorganisms to become unidentifiable (*e.g.* unculturable, unrecognizable, or no longer intact). Analyte particles may also fail to be adequately prepared for enumeration (*e.g.* inadequate staining of microorganisms, overlap or insufficient growth of plaques and colonies, or poor culturability). Many of these losses are unpreventable and unquantifiable, and can only be addressed by evaluating the analytical recovery of the enumeration method.

2.2.3 Sub-sampling Errors

When a sample is processed to a specific stage in an enumeration procedure and then subdivided for further processing (*i.e.* partial sample analysis), the subsample will contain a random number of the analyte particles (even if the sample is well-mixed). Sub-sampling from a suspension (or a centrifuge pellet) that is not well-mixed is a non-random error because it is not possible to predict or model what fraction of the particles were collected in the subsample. The subsample, in such cases, may not be representative of the entire sample and would therefore also not be representative of the source from which the sample was collected. In contrast, it is relatively simple to model sub-sampling errors for discrete particles in well-mixed samples as a function of the fraction of sample that was fully enumerated (Schmidt *et al.*, 2010b; Section 4.2.5). If multiple subsamples are enumerated, then modelling and inference are complicated for two reasons: the quantities of particles in the subsamples are dependent and the analytical errors in enumerating each subsample may be different. The dependence among repeated samples drawn from a source with finite volume (which is analogous to repeated subsamples from a sample) is addressed in Section 4.2.3.

2.2.4 Counting Errors

Counting error describes the difference between the quantity of enumerable particles in a prepared sample and the number of particles observed (Emelko *et al.*, 2010a). Failure to count an observable particle (due to inadvertent under-counting or a false-negative identification error) is a type of loss. Over-counting and false-positive observations, on the other hand, result in observations that do not correspond to actual particles in the sample. Counting errors contribute to the overall analytical error in an enumeration method (as shown in Figure 2.2) and are implicitly included in recovery estimates because they are generally not evaluated and modelled separately from other analytical errors (*e.g.* processed samples are generally not enumerated more than once). Random counting errors can be included in models for analytical error (and the variability in analytical recovery among samples) while non-random errors (*i.e.* substantial mistakes indicating poor control) cannot. Incorporating false-positive observations into analytical error models is particularly challenging because they may depend on the sample matrix or skill of the analyst and, unlike the other counting errors, may not be proportional to the quantity of analyte particles in the sample (Emelko *et al.,* 2010a). For example, it has been demonstrated that some types of algae can result in false-positive enumerations of *Cryptosporidium* oocysts or *Giardia* cysts (Clancy *et al.*, 1994). A model in which analytical recovery can exceed 100% due to random counting errors is discussed in Section 4.2.4. Counting errors can be demonstrated by obtaining different counts from repeated enumerations of a single processed sample, but such results do not indicate whether the counting errors increase or decrease counts and equal counts are not necessarily correct counts (*e.g.* if the same counting error is made repeatedly).

2.2.5 Case Study: Analytical Errors in *Cryptosporidium* and *Giardia* Enumeration

The methods that have been used to enumerate *Cryptosporidium* oocysts and *Giardia* cysts in environmental samples are prone to many types of errors because many steps are required to concentrate and purify these samples prior to enumeration. Finding (oo)cysts in various types of water is often compared to the proverbial process of 'finding a needle in a haystack' because one (oo)cyst in a large volume of water can be important and its observation can easily be obscured by large quantities of other microorganisms or suspended particulate. Samples of 100 L or more are common when enumerating (oo)cysts in raw and finished drinking waters, and it is necessary to concentrate the sample so that the microbes and particulate are present in a smaller volume. For example Envirochek® and Envirochek® HV sampling capsules (Pall Life Sciences, Port Washington, NY) and Filta-Max® foam filters (IDEXX Laboratories, Inc., Westbrook, ME) are commonly used in compliance with Method 1623 (USEPA, 2005b) to concentrate environmental samples. Subsequently, the particulate is eluted from these filters and the eluate is further concentrated by centrifugation (after which much of the supernatant is discarded). The result is a small vial that ideally contains most of the microorganisms and particulate from the original large-volume sample. Purification is then needed to separate the (oo)cysts from the other microorganisms and particulate in the sample. In the Information Collection Rule method (USEPA, 1996), purification consisted of centrifugation in Percoll-sucrose density gradients: dense sediments would settle during centrifugation while (oo)cysts would remain in the upper layers of the immiscible fluid. Method 1623 (USEPA, 2005b) uses immunomagnetic separation (IMS) to purify the sample. In IMS, magnetic beads that attach to (oo)cyst proteins are added to the sample, the beads and attached oocysts are separated from the other particulate, and then the (oo)cysts are dissociated from the beads. Immunofluorescent stains are used to aid microscopic identification and enumeration of (oo)cysts, and additional stains and types of microscopy are used to examine the morphology of presumptive (oo)cysts.

All of the steps in these complicated enumeration methods are prone to analytical errors. It is possible for (oo)cysts to be lost on any apparatus with which the sample has contact (*e.g.* attachment to, or residual sample retained in, the apparatus) or to be inadvertently discarded (*e.g.* in filtrate or supernatant). (Oo)cysts may also fail to be observed due to degradation during sample processing or inadequate staining, or because they are obscured by other matter on the slide. There can also be subsampling errors at any time that a sample is subdivided for further processing (*e.g.* following centrifugation) and there can be errors in counting the number of enumerable (oo)cysts on a prepared microscope slide. Tables 2.1 and 2.2 summarize the methodology and associated possible errors (except for degradation) for two methods that have been used to enumerate *Cryptosporidium* oocysts and *Giardia* cysts: the ICR method (USEPA, 1996; Atherholt and Korn, 1999) and Method 1623 (USEPA, 2005b). Young and Komisar (1999) provided a similar assessment of possible analytical errors in a schematic diagram of the ICR method. The extent to which each of the listed possible errors contributes to the poor and highly variable analytical recovery that is commonly obtained using these methods is largely unknown. Preparing lists of all possible errors in an enumeration method is a helpful tool in quality control and method development because it enables a systematic approach to identify and evaluate analytical errors that may contribute to poor analytical recovery.

Table 2.1: Possible sub-sampling and analytical errors in the Information Collection Rule oo(cyst) enumeration method

Step	**Method Details** (USEPA, 1996; Atherholt & Korn, 1999)	**Possible Errors**
Sample Collection	100 L of untreated water or 1,000 L of treated water are collected	- losses - attachment/residuals in sample collection equipment
Filtration	1 μm nominal porosity yarn-wound polypropylene cartridge filter is used	- losses - attachment to filter apparatus (preceding the filter) - losses - (oo)cysts in filtrate
Elution	filter fibres are washed	- losses - (oo)cysts not eluted from the filter fibres - losses - attachment/residuals in eluate collection apparatus
Concentration	eluate is concentrated by centrifugation and the “packed-pellet volume” obtained from the initial centrifugation step is recorded	- losses - (oo)cysts in aspirated supernatant - losses - attachment/residuals in centrifuge vials
Purification (Flotation)	resuspended pellet is purified by Percoll-sucrose centrifugation (only 0.5 mL of the pellet is used if the packed pellet volume is greater than 0.5 mL)	- sub-sampling error - oo(cyst) dispersion in resuspended pellet - losses - attachment/residuals in sampling equipment
Reconcentration	the top layer, the interface, and the upper portion of the lower layer of the gradient are collected and the material is reconcentrated by centrifugation	- losses - (oo)cysts not collected from lower layer of gradient
Resuspension	supernatant is used to resuspend the resulting pellet of particulate material to a final volume of 5 mL	- losses - (oo)cysts in unused supernatant
Membrane Filtration	1 mL aliquots are applied to membrane filters with 0.2 μm pore-size	- sub-sampling error - oo(cyst) dispersion in resuspended pellet - losses - attachment/residuals in sampling/filtration equipment - losses - (oo)cysts in filtrate
Slide Preparation	filters are prepared by labelling with monoclonal antibody and fluorescent dye-tagged secondary antibody, washing, dehydrating, clearing, and mounting on microscope slides	- losses - insufficiently stained (oo)cysts - losses - obscured (oo)cysts
Enumeration	numbers of cysts and oocysts, with and without certain internal structural features, are counted by the analyst	- counting and identification errors

Table 2.2: Possible sub-sampling and analytical errors in Method 1623

Step	**Method Details** (USEPA, 2005b)	**Possible Errors**
Sample Collection	bulk water sample is collected in a field carboy, and possibly transferred to a laboratory carboy	- losses - attachment/residuals in sample collection equipment
Filtration	sample is filtered using an approved filter capsule or foam filter	- losses - attachment to filter apparatus (preceding the filter) - losses - (oo)cysts in filtrate
Elution	particulate is eluted using the appropriate methodology for the filter that was used	- losses - (oo)cysts not eluted from the filter - losses - attachment/residuals in eluate collection apparatus
Concentration	eluate is concentrated through centrifugation, the pellet volume is recorded, and the pellet is resuspended	- losses - (oo)cysts in aspirated supernatant - losses - attachment/residuals in centrifuge vials
Purification (IMS)	entire resuspended pellet is purified if packed pellet volume is < 0.5 mL; otherwise, one or more aliquots (equivalent to 0.5 mL of packed pellet each) are purified separately	- sub-sampling error - oo(cyst) dispersion in resuspended pellet - losses - attachment/residuals in sampling equipment - losses - (oo)cysts in poured/aspirated supernatant
Dissociation	(oo)cysts are separated from magnetic beads by acid rinses and the sample is applied to one or two well slides	- losses - (oo)cysts retained in the microcentrifuge tube
Slide Preparation	well slide is stained and washed, and mounting medium and a cover slip are added	- losses - insufficiently stained (oo)cysts - losses - (oo)cysts lost in slide washing - losses - obscured (oo)cysts
Enumeration	slide is examined using FITC, DAPI fluorescence, and DIC examination	- counting and identification errors

2.3 DEFINITION OF ANALYTICAL RECOVERY

In addition to the term ‘recovery’ having diverse technical meanings in numerous fields, the term ‘analytical recovery’ has also been defined in several different ways and has many alternative names (*e.g.* ‘recovery’, ‘recovery efficiency’, ‘percent recovery’, ‘probability of recovery’, ‘recovery rate’). These terms, which can sometimes have subtly different interpretations, have often been used interchangeably and without adequate definition. In general, analytical recovery describes some relationship between the quantity of analyte particles that are observed and the quantity that are actually present (*e.g.* “the portion of microorganisms identified by a particular enumeration method with respect to the number that were actually initially present in the water sample” – Petterson *et al.*, 2007). The term ‘recovery’, however, has also been used as a synonym for detection (*e.g.* Clancy *et al.*, 1994) because it can appropriately be said that detected particles have been recovered from the sample. Accordingly, it has been said that “recovery is related directly to the concentration of cysts and oocysts in the sample” (Clancy *et al.*, 1994). Such statements, based on ambiguous definitions of recovery, are misleading. The statement is intended to mean that detection of (oo)cysts in a sample is more probable when more (oo)cysts are present because the methodology is susceptible to substantial losses that may lead to false-negative enumerations when few (oo)cysts are present. Taken in a different context, the statement implies that the (oo)cyst concentration in the samples affects the relative magnitude of losses or counting errors in the methodology, which is not supported by the associated results (because all samples were prepared with the same concentration). This section discusses the conventional definitions of analytical recovery and proposes a more general definition.

Analytical recovery has typically been regarded as the fraction of the particles in a sample that are observed, and has occasionally been defined as the probability that each particle in a sample will be observed. The analytical recovery of a particular method is quantified by enumerating samples that are seeded with known quantities of analyte particles (Section 4.1). The fraction of these seeded particles that are observed in a sample (*i.e.* the number of observed particles divided by the presumed number of seeded particles), or the associated ‘percent recovery’ (*e.g.* USEPA, 2005b), is then used to describe analytical recovery. These fractions, however, are necessarily a function of the number of particles seeded into the sample (*e.g.* it is not possible to observe 50% of one seeded particle). What is of interest in most applications is not how well precisely known quantities of seeded particles have been enumerated, but how well unknown quantities of indigenous particles would be enumerated. In this sense, analytical recovery is not measured directly, but is estimated from the fraction of seeded particles that are observed. Accordingly, the term ‘recovery estimate’ is used throughout this thesis to describe the fraction of seeded particles that are observed: ‘recovery measurement’ is a misleading term because the analytical recovery of indigenous particles cannot be measured directly unless the quantity actually present is known *a priori*.

The majority of publications that address quantification of analytical recovery (*e.g.* to demonstrate the proficiency of a particular method, to compare methods, or to show compliance with validation criteria) regard the fraction of seeded particles that are observed as though it is an exact measurement of analytical recovery. Seeding errors (Section 4.1.1), however, cause the computed fractions of seeded particles that are observed to be imprecise, and the discrete nature of particles causes the fraction of seeded particles that are observed to have discrete (and random) values that depend upon how many particles are observed. Accordingly, the results from several recovery experiments are only directly comparable if the number of seeded particles (and possible variability therein) are the same. For this reason, reported recovery values (or summary statistics thereof) have minimal scientific value unless they are also accompanied with details about the quantities of seeded particles (and possible errors therein). The term ‘recovery data’ is used herein to describe the set of raw counts and seed doses (along with a discussion of how the seed doses were prepared or how precise they are) rather than the comparatively uninformative fractions of seeded particles that were observed.

Probabilistic modellers have used a more general definition of analytical recovery that is not conditional upon the quantity of particles that were actually in the sample, because this quantity is usually unknown in unseeded (*e.g.* environmental) samples. Accordingly, recovery is not regarded as an empirical measurement of losses but as an intrinsic property of the method that determines how many of the present particles are likely to be observed. Many probabilistic models, especially those that use a binomial distribution to describe analytical error (*e.g.* Sections 4.2.2, 4.2.3, 4.2.6, and 4.2.7), have treated analytical recovery as a probability (*e.g.* Nahrstedt and Gimbel, 1996; Parkhurst and Stern, 1998; Teunis *et al.*, 1999; Teunis and Havelaar, 1999; Crainiceanu *et al.*, 2003; Petterson *et al.*, 2007; Schmidt *et al.*, 2010a; Emelko *et al.*, 2010a). Defining analytical recovery as a probability is acceptable to describe losses, but not to model counting errors that can potentially lead to counts exceeding 100% of the seeded particles (as described in Section 2.2.4). A more general definition of analytical recovery can be obtained by regarding it as a rate. This encompasses the probability-based definition (because probability can be regarded as the long-run rate of successes in repeated success-failure trials) and also enables recovery to exceed 100%. Accordingly, analytical recovery is defined herein as “the capacity of the analyst to successfully count each microorganism or particle of interest in a sample using a specific enumeration method” (Schmidt *et al.*, 2010a).

Chapter 3
Introduction to Probabilistic Modelling and Quantitative Analysis Approaches

This chapter provides a brief introduction to the statistical tools that are used throughout this thesis. The general research approach employed in this thesis consists of three steps. The first step includes careful consideration of all possible causes of random variability in enumeration data and development of general probabilistic models that describe each source of variability using basic mathematical assumptions and appropriate distributions. These models are then used to quantify uncertainty in parameters of interest (*e.g.* concentration, analytical recovery) given the available data. Finally, the models are also used to explore relative contributions of different sources of variability to the overall variability in enumeration data and to compare strategies that may enable collection of more precise estimates of parameters such as analytical recovery and concentration. Section 3.1 addresses probabilistic modelling, and Section 3.2 addresses general quantitative analysis approaches that can be used to interpret available data. Section 3.3 describes ways in which probabilistic models can be used to simulate data or to quantify the effects of model parameters upon enumeration data.

3.1 PROBABILISTIC MODELLING

As described in Section 2.1, there can be many reasons for which the count of microorganisms or discrete particles in a sample might vary. A systematic approach is needed to evaluate the errors (or sources of variability) in enumeration data. These errors may include (1) natural variation in the parameter being quantified (*e.g.* concentration heterogeneity), (2) recorded differences in sample characteristics or the enumeration process (*e.g.* different sample volumes among samples), (3) random errors associated with the sample collection and enumeration process (*e.g.* randomness of the number of particles contained in a sample), or (4) non-random mistakes by the analyst. It is assumed throughout this thesis that enumeration methods are used with sufficient quality control that non-random errors can be ignored and that differences between samples (*e.g.* the processed sample volume or the methodology used) are accurately recorded. All remaining variability among enumeration data can then be attributed to random measurement errors or to natural variation in the parameter being quantified. It is essential to understand the relative contribution of each source of variability to the data because incorrect quantitative analysis of enumeration data can yield biased results (*e.g.* “Is the observed variability due to variability in concentration or to measurement error?”).

The models that are presented in this thesis describe random variation among enumeration data using a hierarchical sequence of distributions that represents the sequence of errors that make enumeration data variable. The distributions are chosen and arranged in accordance with simple mathematical assumptions rather than case-specific fitting and correlation. Understanding the processes that contribute variability to data conveniently enables the development of generally applicable probabilistic models without the need for extensive replication that would enable construction of case-specific fitted models. Regression models, for example, are very useful investigative tools to evaluate possible covariates or the relative contributions of different sources of variability, but they require large amounts of data and the results are often case-specific and do not necessarily explain the underlying true relationships between different factors. This section describes important statistical terminology that is used throughout the thesis, provides a summary of the distributions that are used in this thesis, and describes two approaches that can be used to obtain point estimates of unknown model parameters from available data.

3.1.1 Important Concepts

Accuracy and precision

Measurement error in any method that is used to obtain data compromises the precision, and possibly also the accuracy of results. Accuracy describes the proximity of a datum (or the mean of fully replicated data) to the true value of the parameter that is being measured, while precision concerns the consistency of (or absence of variability among) fully replicated measurements. Individual measurements may be inaccurate due to high variability (*e.g.* some measurements may fall near the true value while others may be quite different) and a measurement process may be inaccurate due to bias (*i.e.* measurements that are systematically different from the true value). Reducing measurement errors (*i.e.* improving precision and removing bias) is essential to obtain more accurate data. Proper understanding of measurement errors and use of suitable quantitative analysis approaches are essential to make appropriate decisions on the basis of available data.

Bias

Bias describes a systematic difference between a statistic (*i.e.* a function of measured data) and the true value associated with the population it represents. For example, if estimated concentrations are systematically (*i.e.* on average) lower than the true concentrations they represent, then the enumeration methodology and associated data are biased. Likewise, if the variability among concentration estimates is systematically greater than the variability in concentration in the source from which the samples were drawn, then the variability among concentration estimates is a biased estimate of variability in concentration. Bias can result from improper sample collection (*i.e.* strategically or inadvertently gathering samples that are not collectively representative of the population that they are intended to represent), measurement errors (*i.e.* obtaining 'measurements' that are systematically different from their associated true values), or incorrect data analysis (*e.g.* using statistical methods that are based on faulty assumptions or that disregard measurement errors in the data). Bias increases the possibility of making incorrect decisions (*e.g.* drawing conclusions that are not grounded in reality even though they appear to be supported by the available data).

Replication

In general, replication is the process of repeating measurements under similar conditions to yield information about the precision of the measurement process. Repeated measurements are fully replicated if they can be regarded as independent and identically distributed measurements. The resulting variation is attributed exclusively to random measurement error rather than variability in the parameter being measured or other controllable factors. For example, fully replicated concentration estimates would require identically collected and processed water samples from a homogeneous (*i.e.* well-mixed) source. Relative to more complicated sets of data, this type of data can easily be analyzed by many conventional statistical analysis techniques (especially if the data are approximately normally distributed or sufficiently abundant to enable transformation) because the data are homoscedastic (*i.e.* they have equal measurement error). Many elementary statistical tools (*e.g.* hypothesis tests and confidence intervals based on the t, chi-squared, or F distribution, and analysis of variance) assume that data (or the associated measurement errors) are approximately normally distributed with constant variance.

Throughout this thesis, replication is more broadly regarded as collecting repeated enumeration data that are believed to be representative of a constant concentration, or collecting repeated recovery estimates that are believed to represent only random variation in analytical recovery among samples. For example, replicate concentration estimates are obtained by processing repeated samples from the same source, but these may have different sample volumes or be processed using different methodology. Concentration estimates can only be regarded as replicates if it is assumed that there is no temporal or spatial variability in concentration in the source from which the samples were drawn. Replicate recovery estimates are obtained from samples with equivalent water quality and identical sample processing and enumeration methodology, but may contain different quantities of the microorganisms or particles being enumerated (if it is assumed that recovery is independent of the quantity of seeded microorganisms or discrete particles). Recovery estimates can only be regarded as replicates if it is assumed that there are no controllable differences in sample characteristics or methodology that may impact analytical recovery.

Statistical control

Statistical control is a characteristic of the method used to obtain data that indicates whether variability in the data arises from random or non-random errors. “An experimental procedure is said to be in a state of statistical control when the observations to which it gives rise, under what are assumed to be ‘essentially the same conditions’, fluctuate in a random manner and are free from trends and non-random shifts in magnitude” (Eisenhart and Wilson, 1943). Variability is random when its causes are immeasurable (or at least when the variability is adequately modelled using the statistical concepts of random variables and distributions). In contrast, it is at least partially deterministic if outcomes can be predicted using covariate data (or if repeated measurements seem to follow a non-random pattern). If a probabilistic model that regards measurements as random variables (with variability that is modelled by a distribution) adequately describes the variability in successive measurements, then the method is in a state of statistical control. “Unless a sampling procedure – and in a sense all processes leading to observations are sampling procedures – is in a state of statistical control, it is not possible to make valid inferences about the 'population' which the observations are supposed to represent" (Eisenhart and Wilson, 1943). All statistical inference depends upon statistical control and models that appropriately describe randomness.

Variability and Uncertainty

"Variability corresponds to the changing nature of a variable that changes over time, over space, or among samples, or to variation among data" (Schmidt and Emelko, 2010a): for example, concentration varies over time and concentration estimates vary. In contrast, "uncertainty represents an imperfect state of knowledge about a parameter or a model" (Schmidt and Emelko, 2010a): for example, inaccuracy of a concentration estimate, or inaccuracy of a model (or its parameters) that is used to describe variability in concentration. The distinction between variability and uncertainty is not always clear. For example, the distribution representing variability in analytical recovery for a particular enumeration method can also be used to model uncertainty in the recovery of indigenous analyte particles in a specific sample. Knowledge of the difference between variability and uncertainty is essential to develop appropriate conceptual models (*e.g.* quantitative microbial risk assessment models) and associated statistical tools.

3.1.2 Utilized Distributions

The following simple distributions are used in this thesis because of their mathematical properties. In general, discrete distributions are used to describe counts of analyte particles (at various stages throughout the sampling and enumeration process) and continuous distributions are used to describe parameters such as concentration or analytical recovery. The symbols for the random variable and parameters that are used in this section are not the same as what is used elsewhere in the thesis because many of these distributions are used in several different ways and each may have different symbols for the parameters and random variables.

The binomial distribution

The binomial distribution is a discrete distribution with a random variable (X) representing the number of successes in a specified number (n) of independent (Bernoulli) trials with equal probability (p) of success (Evans *et al.*, 2000). It is used to represent two types of Bernoulli processes herein: (1) the number of analyte particles in a sample that are successfully observed using a particular enumeration method when each particle has an equal probability of being observed (*e.g.* Sections 4.2.2, 5.1.1, 5.1.3), and (2) the number of analyte particles contained in a specified fraction of a well-mixed sample that contains a specific number of such particles (*e.g.* Sections 4.2.5, 5.1.3). The probability mass function, mean, and variance of the binomial distribution are presented in Table 3.1.

Table 3.1: Probability functions and properties of utilized distributions

Discrete Distribution	**Probability Mass Function**	**Constraints**	**Mean**	**Variance**
Binomial	$f_X(x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$	$\{x\in I \mid 0\le x\le n\}$ $\{n\in I \mid n> x\}$ $\{p\in R \mid 0< p<1\}$	np	$np(1 - p)$
Poisson	$f_X(x)=\frac{e^{-\lambda t}(\lambda t)^x}{x!}$	$\{x\in I \mid x\ge 0\}$ $\{\lambda\in R \mid \lambda> 0\}$ $\{t\in R \mid t\ge 0\}$	λt	λt
Negative binomial (Bernoulli process formulation)	$f_X(x)=\frac{(x+k-1)!}{x!(k-1)!}p^k(1-p)^x$	$\{x\in I \mid x\ge 0\}$ $\{k\in I \mid k\ge 1\}$ $\{p\in R \mid 0< p<1\}$	$k\frac{(1-p)}{p}$	$k\frac{(1-p)}{p^2}$
Negative binomial (alternative formulation)	$f_X(x)=\frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)}\left(\frac{1}{\beta+1}\right)^{\alpha}\left(\frac{\beta}{\beta+1}\right)^{x}$	$\{x\in I \mid x\ge 0\}$ $\{\alpha\in R \mid \alpha> 0\}$ $\{\beta\in R \mid \beta> 0\}$	$\alpha\beta$	$\alpha\beta(\beta+1)$
Continuous Distribution	**Probability Density Function**	**Constraints**	**Mean**	**Variance**
Beta	$f_X(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$	$\{x\in R \mid 0< x<1\}$ $\{a\in R \mid a> 0\}$ $\{b\in R \mid b> 0\}$	$\frac{a}{a+b}$	$\frac{ab}{((a+b)^2(a+b+1))}$
Gamma	$f_X(x)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}$	$\{x\in R \mid x> 0\}$ $\{\alpha\in R \mid \alpha> 0\}$ $\{\beta\in R \mid \beta> 0\}$	$\alpha\beta$	$\alpha\beta^2$
Normal	$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\{x\in R\}$ $\{\mu\in R\}$ $\{\sigma\in R \mid \sigma> 0\}$	μ	σ^2

The Poisson distribution

The Poisson distribution is a discrete distribution with a random variable (X) that represents the number of observations in a sample of size (t) in which observations occur randomly and independently at rate (λ). It is also the limiting form of the binomial distribution when n is large and p is small (Evans *et al.*, 2000). It is used herein to represent the number of analyte particles in a sample of specified volume that is drawn (independently of any other samples) from a source with homogeneous concentration (*e.g.* Sections 4.2.3, 4.2.4, 5.1.1, 5.1.2). This model assumes that the analyte particles are randomly dispersed throughout the source (*e.g.* not clumped). The probability mass function, mean, and variance of the Poisson distribution are presented in Table 3.1. The mean and variance of the Poisson distribution are necessarily equal and the distribution can be defined by a single parameter (λt). Another property of the Poisson distribution that is used throughout this thesis (*e.g.* Sections 4.2.3, 4.2.5, 5.1.1, 5.1.3) is that a random variable Y, which is binomially distributed with probability of success θ and a number of Bernoulli trials (X) that is itself Poisson-distributed with mean λt, is Poisson-distributed with mean $\lambda t\theta$.

The negative binomial distribution

The negative binomial distribution is a discrete distribution with a random variable (X) that can be defined in several different ways. It is commonly used in association with a Bernoulli process to represent the number of failures after which the k^{th} success occurs (with constant probability of success p). This is also known as the Pascal distribution (Evans *et al.*, 2000). The probability mass function, mean, and variance of this type of negative binomial distribution are presented in Table 3.1.

A negative binomial distribution will also result from a situation in which X is Poisson-distributed with a mean that is itself gamma-distributed (with parameters α,β as described below) (Fisher, 1941; Emelko *et al.*, 2010b). This alternative formulation of the negative binomial distribution enables the parameter k (now α) to be a positive real number rather than a positive integer. The probability mass function, mean, and variance, of this formulation of the negative binomial distribution are also presented in Table 3.1. This formulation of the negative binomial distribution is used when the number of observed analyte particles is Poisson-distributed (with mean equal to the product of sample volume, source concentration, and analytical recovery) and the analytical recovery of the method is assumed to be gamma-distributed (*e.g.* Sections 4.2.4, 5.1.2).

The negative binomial model is commonly considered to fit data that are over-dispersed relative to the Poisson distribution (*e.g.* Pipes *et al.*, 1977; El-Shaarawi *et al.*, 1981; Christian and Pipes, 1983; Haas and Heller, 1986; Gale *et al.*, 1997) because its variance is necessarily greater than its mean. The negative binomial distribution converges asymptotically upon the Poisson distribution (Evans *et al.*, 2000) as p approaches unity (with large values of k) or as β approaches zero (with large values of α).

The beta distribution

The beta distribution is a continuous distribution with a random variable (X) that has two shape parameters (a,b) and that is typically restricted to the interval $0<X<1$ (Evans *et al.*, 2000). The probability density function, mean, and variance of the beta distribution are shown in Table 3.1. The standard continuous uniform distribution is a special case of the beta distribution in which a and b each have a value of unity.

Use of the beta distribution in the models presented in this thesis is not the natural result of simple mathematical assumptions: it is chosen because of its mathematical properties. When the random variable is a probability (*e.g.* recovery in Sections 4.2.2, 4.2.3, 5.1.1), it is convenient to use a distribution that is bounded on the interval [0,1]. The beta distribution is also unimodal (unless both parameters are less than or equal to unity) and can be skewed (positively skewed if the mean is less than 0.5, symmetrical if the mean is 0.5, and negatively skewed if the mean is greater than 0.5). The beta distribution is also the conjugate prior of the binomial distribution (Ang and Tang, 1975) as discussed in Section 3.2.2.

The gamma distribution

The gamma distribution is a continuous distribution with a random variable (X) that is restricted to positive values and that has two parameters (shape parameter α and scale parameter β). The probability density function, mean, and variance of the gamma distribution are shown in Table 3.1. The gamma distribution has several special cases (Evans *et al*., 2000): the exponential distribution (the waiting time until the first event in a Poisson process with rate $1/\beta$), the Erlang distribution (the waiting time until the α^{th} event in a Poisson process with rate $1/\beta$), and the chi-squared distribution (the sampling distribution of the sample variance when data are independent and identically normally distributed).

Like the use of the beta distribution, use of the gamma distribution in the models presented in this thesis is not the natural result of simple mathematical assumptions: it is chosen because of its mathematical properties. Many environmental continuous random variables are positively skewed and assume only positive values (like the gamma distribution). The gamma distribution is also the conjugate prior of the Poisson distribution (Ang and Tang, 1975) as discussed in Section 3.2.2 and is a convenient distribution to describe variability in the mean of a Poisson distribution (a negative binomial model results as described above). The gamma distribution is used herein to describe variability in analytical recovery when recovery can exceed 100% (*e.g.* Sections 4.2.4, 5.1.2) and to describe temporal variability in concentration (Section 6.1.1).

The normal distribution

The normal distribution is a continuous distribution with a random variable (X) that can assume any real value (*i.e.* $-\infty < X < \infty$) and that has two parameters (location parameter μ and scale parameter σ) (Evans *et al*., 2000). The normal distribution is sometimes inappropriate to model random variables that are restricted to positive values (*e.g.* recovery, concentration), but can sometimes be used to approximate other types of distributions (*e.g.* a binomial or Poisson distribution with a high mean, a negative binomial distribution with a high value of k or α, or a gamma distribution with a high value of α). The assumption that data are approximately normally distributed is common in conventional statistical tools (*e.g.* t-tests, chi-squared tests, F-tests, analysis of variance for regression models). The normal distribution also arises from the central limit theorem: the sum of a large number of independent deviates drawn from any distribution with finite variance is approximately normally distributed. In accordance with the central limit theorem, the normal distribution is used in this thesis to model the sampling distribution of the mean when there are relatively large numbers of data (or the data are approximately normally distributed) and the variance is known (*e.g.* Sections 8.1.4, 8.2.2).

3.1.3 Parameter Estimation

Once a distribution has been chosen to model a particular set of data (because it follows naturally from simple mathematical assumptions, has convenient properties, or has good case-specific fit), it is necessary to estimate the unknown parameters of the distribution. Two commonly used point-estimation approaches are addressed in this thesis: the method of moments and maximum likelihood estimation. Such point estimates are commonly reported without appropriate consideration of uncertainty. As described in Section 3.2.2, Bayes' theorem (which makes use of the likelihood function) can be used to provide a quantitative description of uncertainty in estimated parameters.

Method of moments

The method of moments is a simple parameter estimation technique that equates the moments of the distribution to those of the sample when the data are independent and identically distributed. For example, the mean is the first moment (or expected value) of the random variable X ($\mu = \mathrm{E}[X]$) and the variance is the second moment about the mean ($\sigma^2 = \mathrm{E}[(X - \mu)^2]$). The number of independent moments required to fit a distribution by the method of moments is equal to the number of parameters in the distribution. For example, when beta or gamma distributions are fitted to recovery data using the method of moments (Section 4.4.1), the sample mean and sample variance are used because the beta and gamma distributions each have two parameters. The method of moments can yield biased results if measurement errors in the data are ignored (*e.g.* Section 4.4.1) and can sometimes yield infeasible parameter values (*e.g.* Section 8.1.5).

Maximum likelihood

Maximum likelihood (Fisher, 1925) is a more complex parameter estimation technique than the method of moments that can be used to estimate parameters of a distribution with full consideration of the probabilistic process by which the data are generated. The likelihood function (Equation 3.1) is defined as the joint probability of the data (the vector X) conditional upon specific, feasible values of the parameters being estimated (the vector θ). If the data are independent, then the joint probability of the data is the product of the probability (conditional upon specified values of the parameters) associated with each of the r individual data (Equation 3.2). This method does not necessarily assume that the data are all identically distributed: for example, repeated concentration measurements from a homogeneous source that have different sample volumes or that were processed using different methodology can be accommodated. If the data are not independent, then the relationship between the data must be addressed in the likelihood function (*e.g.* Section 6.1.1).

$$L(\theta) = f_X(X \mid \theta) \tag{3.1}$$

$$L(\theta) = \prod_{i=1}^{r} f_{X_i}(x_i \mid \theta) \tag{3.2}$$

Once the likelihood function has been defined, maximum likelihood estimates are obtained by finding the global maximum of the function within the feasible parameter space. Optimization is complicated except in relatively simple cases because the global maximum often cannot be calculated explicitly (*e.g.* Appendix A) and the likelihood itself cannot always be calculated explicitly (*e.g.* Section 4.4.1). Broad likelihood functions (*i.e.* those with likelihood values near the maximum over a wide region of the parameter space) indicate high uncertainty in the estimated parameter values because many possible parameter values have comparably high likelihoods. Maximum likelihood estimation is used in this thesis to obtain point estimates for the parameters of beta and gamma distributions that describe random variability in recovery among samples (Section 4.4.1)

3.2 QUANTIFICATION OF UNCERTAINTY IN PARAMETER ESTIMATES AND TESTS OF HYPOTHESES

Data are generally collected to measure or model some sort of process (*e.g.* the variability in recovery among replicates, the concentration of microorganisms in a water supply, the removal or inactivation of analyte particles by a treatment process, or the probability of infection due to consumption of water that may contain pathogens) or to make some sort of decision (*e.g.* "Does a laboratory meet validation criteria for an enumeration method?", "Is one enumeration method superior to another?", or "Given measured raw water quality, is additional treatment required to comply with a regulation?"). Models and decisions, however, are typically uncertain because they are based upon limited data (*e.g.* it is not possible to enumerate pathogens in every portion of a drinking water supply while supplying drinking water). Moreover, the data upon which these models and decisions are based may be inaccurate and/or imprecise due to measurement errors. Therefore, analyses and decisions that are based exclusively upon raw data (*i.e.* without consideration of measurement errors and uncertainty) are prone to bias (which can lead to wrong decisions or misleading models). Proper consideration of uncertainty is needed to make appropriately informed decisions and to develop useful models. This section summarizes the strategies that are used in this thesis to quantify uncertainty in concentration and analytical recovery estimates, model parameters, and decisions. Specifically, it summarizes types of hypothesis tests (and their corresponding confidence intervals) that are used herein and provides an introduction to Bayesian statistics.

3.2.1 Hypothesis Testing and Confidence Intervals

Hypothesis testing is a framework by which decisions can be made about an imperfectly known parameter given the available data, while confidence intervals are a quantitative measure of uncertainty in the parameter. In addition to ensuring that critical assumptions are met (*e.g.* many common hypothesis tests and confidence intervals assume that the data are independent and normally distributed and have constant variance), care must be taken when conducting hypothesis tests or calculating confidence intervals based on data that are prone to measurement errors (*e.g.* Sections 4.4.5 and 5.5). Detailed discussion of the theory and methodology for hypothesis tests (and associated confidence intervals) is beyond the scope of this thesis. The following hypothesis tests (or confidence intervals) and their associated assumptions are presented because they are specifically addressed in this thesis.

Hypothesis tests and confidence intervals for a single mean

There are two types of common hypothesis tests (or confidence intervals) for a single mean. The first arises from the central limit theorem, which states that the sum of any number of normally distributed values (with constant variance) is normally distributed and that the sum of a large number of non-normally distributed values (with finite population variance) is approximately normally distributed. Accordingly, the sampling distribution of the mean ($\overline{X}$) is also normally distributed (Equation 3.3) if the data are normally distributed or if there are a large number (*n*) of data that follow a distribution with finite population variance. This sampling distribution can be used in hypothesis tests or confidence intervals for a single mean when the population variance (σ) is precisely known. It is used to quantify the power of hypothesis tests on a single mean in Section 4.4.5 and to calculate probability intervals for the sample mean of recovery and concentration estimates in Sections 8.1.4 and 8.2.2 respectively. The hypothesis testing procedure and confidence interval associated with this scenario are not considered herein because the scenario in which the population mean is of interest but the population variance is known is uncommon.

$$f_{\overline{X}}(\overline{x}) \sim NORMAL\left(\mu, \sigma/\sqrt{n}\right) \tag{3.3}$$

The second common hypothesis test (or confidence interval) for a single mean is specific to independent, normally distributed data when the population variance is unknown. In this case, the sampling distribution of the mean is normally distributed as per Equation 3.3 but the standard error ($\sigma/\sqrt{n}$) is unknown. The sampling distribution of the mean is related to a t distribution with *n*-1 degrees of freedom as shown in Equation 3.4. Use of these statistics (or similar statistics based on the t distribution that are used to test the difference between two means) for data that are only approximately normally distributed will sometimes yield a valid approximation (Box, 1953). Equation 3.5 shows the confidence interval formula for the population mean that is associated with the statistic shown in Equation 3.4. This statistical inference approach is compared with Bayesian approaches to estimate the population mean of fully replicated (and approximately normally distributed) recovery and concentration estimates in Sections 4.4.5 and 5.5.1 respectively. Specifically, confidence intervals are contrasted with credible intervals and *P*-values are contrasted with posterior probabilities.

$$T_{n-1} = \frac{\overline{X} - \mu_o}{S/\sqrt{n}} \tag{3.4}$$

$$\overline{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \tag{3.5}$$

Other types of confidence intervals and hypothesis tests based on the t distribution are common in statistics (*e.g.* tests on the difference between two means, tests on the coefficients of a linear regression model). In general, these tests assume that the data (or the error term in a regression model) are approximately normally distributed with constant variance.

Hypothesis tests and confidence intervals for a single variance

If data are independent, normally distributed, and have constant variance, then the sampling distribution of the variance is related to a chi-squared distribution with *n*-1 degrees of freedom as shown in Equation 3.6. The associated confidence interval formula is shown in Equation 3.7. This sampling distribution is used to calculate probability intervals for the sample standard deviation of recovery estimates (by taking the square root of the probability interval for the variance) in Section 8.1.4. Statistical inference based on the chi-squared distribution is compared with Bayesian approaches to estimate the population variance of fully replicated (and approximately normally distributed) recovery estimates in Section 4.4.5.

$$\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma_o^2} \tag{3.6}$$

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \tag{3.7}$$

Other types of confidence intervals and hypothesis tests are also based on the chi-squared distribution. The F distribution (which is commonly used to test the difference between two variances or sums of squares, especially in analysis of variance) is derived from the quotient of two chi-squared-distributed variables and also assumes that each set of replicated data is normally distributed with constant variance. In analysis of variance, it is explicitly assumed that the error term in the regression model is normally distributed and has constant variance. Box (1953) has addressed the inaccuracy of F-tests when they are used for data that violate these assumptions.

Fisher’s index of dispersion

Fisher’s index of dispersion (Fisher *et al.*, 1922) is a statistic that is used in a hypothesis test that can be used to evaluate the dispersion of counts in a presumed Poisson process. The test is commonly used (*e.g.* Eisenhart and Wilson, 1943; El-Shaarawi *et al.*, 1981; Margolin, *et al.*, 1981; Christian and Pipes, 1983; Haas and Heller, 1986; Gale *et al.*, 1997) to evaluate the distribution of analyte particles in water (*i.e.* the validity of a Poisson model for fully replicated enumeration data). As described in Section 3.1.2, use of a Poisson model to describe the number of analyte particles in a volume of water drawn from a homogeneous source assumes that the particles are randomly distributed throughout the water. A clumped distribution of particles in the source will cause fully replicated data to be over-dispersed relative to a Poisson model, while evenly distributed particles will cause the data to be under-dispersed. The calculated index of dispersion (Equation 3.8) follows a chi-squared distribution with *n*-1 degrees of freedom if the particles are randomly distributed. This hypothesis test, and other strategies to evaluate the dispersion of analyte particles in water, is addressed in Section 5.2.2.

$$D^2 = \frac{(n-1)S^2}{\bar{X}} \tag{3.8}$$

3.2.2 Bayesian Statistics

Statistical approaches that are based upon Bayes' theorem enable subjective information to be mathematically incorporated into quantitative analysis and they often describe uncertainty in unknown parameters (which have fixed but unknown values) using distributions. This section provides a brief overview of Bayesian statistics, which are used throughout this thesis. The other commonly used statistical approach is that of frequentist statistics (which does not include subjective information or use statements of probability to describe uncertainty in unknown parameters).

Bayes' theorem

Bayes' theorem has two general forms. The most widely recognized form (as represented by Equation 3.9) arises from basic set theory and conditional probability.

$$P(B \mid A) = \frac{P(B) \cdot P(A \mid B)}{P(A)} \tag{3.9}$$

When the theorem is expanded to make probability statements about possible values of the unknown parameters (vector θ) of a probabilistic model given a vector of data (*X*), it takes the form of Equation 3.10.

$$f(\theta \mid X) \propto f(\theta) \cdot f(X \mid \theta) \tag{3.10}$$

In Equation 3.10, $f(\theta|X)$ is the joint posterior probability of the vector of parameters (θ) given the vector of data (*X*), $f(\theta)$ is the prior, and $f(X|\theta)$ is the likelihood (as described in Section 3.1.3).

Priors

The prior is a function that describes the analyst's subjective beliefs about the possible values of the unknown parameters of interest in the model before considering the data. Any unknown parameter that is fully specified in a hierarchical model (*e.g.* a parameter that can theoretically be integrated out in the likelihood function) is a nuisance parameter that should not be included in the prior. If the prior is a distribution (*i.e.* its value is non-negative for all feasible values of θ and its sum/integral has a value of unity), then it is proper. An unbounded uniform prior is an example of an improper prior, because the sum/integral of a constant over infinite or semi-infinite domain is infinite. In some cases, an improper prior can result in an improper posterior (*e.g.* Section 4.4.3). A prior can also be informative or relatively uninformative. A relatively uninformative prior contributes minimal subjective information to the posterior distribution, and the posterior distribution is determined mainly by the available data. An informative prior represents subjective information (arising from previous data or the analyst's beliefs). Even an informative prior may have weak influence on the posterior distribution if the weight of evidence provided by newly collected data is strong. In general, relatively uninformative priors are used throughout this thesis because informative priors are based on subjective information or beliefs and are case-specific. The uniform prior is often used herein because it is often convenient when the posterior is directly proportional to the likelihood for all feasible values of the parameters.

A conjugate prior is a prior that (when paired with data from a specific form of distribution) will result in a posterior of the same form as the prior (Ang and Tang, 1975). For example, a beta-distributed prior for the probability of success in binomially distributed data will result in a beta-distributed posterior (*e.g.* Section 5.3.4). Likewise, a gamma-distributed prior for the mean of Poisson-distributed data will result in a gamma-distributed posterior.

Posterior distributions

A posterior is a distribution (or a density function if it is improper) obtained from Bayes' theorem that describes the (relative) probabilities of possible values of unknown parameters (or possible future values of random variables) given the model, the selected prior, and the available data. There are several types of posterior distributions that can be obtained when a probabilistic model has more than one unknown parameter (joint posterior distributions, conditional posterior distributions, and marginal posterior distributions) and there can also be posterior predictive distributions.

- A joint posterior distribution describes the joint posterior probability of a vector of unknown parameters (*i.e.* the probability function concurrently regards several unknown parameters as random variables). This type of posterior distribution is obtained from Equation 3.10.
- A conditional posterior distribution describes the posterior probability of a single unknown parameter given specific values of all other unknown parameters. This type of posterior distribution is particularly useful in Gibbs sampling as described below.
- A marginal posterior distribution describes the posterior probability of a subset of unknown parameters (*e.g.* a single unknown parameter) across all possible values of the other unknown parameters. For example, marginal posterior distributions for analyte particle concentration (estimated from replicate data) are addressed in Section 5.3.
- A posterior predictive distribution describes the joint posterior probability of a set of random variables that are derived from the unknown parameters in the joint posterior distribution obtained using Bayes' theorem. For example, when Bayes' theorem is used to estimate the probability (p) of getting 'heads' from repeated tosses of an unfair coin, the probability of getting 5 heads in 10 future tosses is a (binomial) random variable that

depends upon the posterior distribution $f_P(p)$. Posterior predictive distributions for concentration are addressed in Section 6.3.

Calculation of posterior distributions is often complicated because the integrating factor that is required to convert the proportionality in Equation 3.10 to an equation often cannot be evaluated explicitly, or the integration or summation that is needed to obtain a marginal posterior distribution from a joint distribution cannot be evaluated explicitly. In some cases, numerical integration can be used to obtain an accurate approximation of the posterior distribution (*e.g.* Chapter 5). In others, integration may be computationally intractable. Markov Chain Monte Carlo techniques (such as Gibbs sampling) are commonly used in such situations.

Gibbs sampling

Gibbs sampling is a relatively simple process that is used to generate a Markov chain of values from a posterior distribution using the conditional posterior distributions of the various unknown parameters. Given sufficient mixing and convergence, the Markov chain of values can be regarded as a sample from the posterior distribution (*i.e.* the relative frequency of values in any range will tend towards its respective posterior probability as the length of the Markov chain is increased). Gibbs sampling is used for many applications in this thesis (*e.g.* Sections 4.4.3, 5.3.2, 5.4.2, and 6.3.1).

Drawing random samples from joint or marginal posterior distributions is often complicated, while drawing random samples from conditional posterior distributions is comparatively simple. It is especially simple if conditional posterior distributions can be arranged in the form of common distributions for which random number generating algorithms are available. Gibbs sampling commences with specified values (which are arbitrary as long as they are feasible) of all but one of the unknown parameters. A value of the remaining unknown parameter is then drawn from its conditional posterior distribution. Each of the unknown parameters (in a repeating sequence) is then updated using the associated conditional posterior distribution and the most recent value of all other unknown parameters. The sequence of generated values is a Markov chain because the new simulated values of a set of parameters (from a full cycle) depends upon the values of the parameters at the outset of the cycle.

It is usually possible (with probability equal to the joint posterior probability of all unknown parameters) for the Markov chain to reach any location in the feasible space of the parameters because the conditional posterior distribution represents all possible values of the unknown parameter (with appropriate probability) that are supported by the available data and specific values of the other unknown parameters,. Consequently, a random sample of elements from an infinitely long Markov chain (generated by Gibbs sampling) is a random sample from the associated posterior distribution.

Mixing and convergence

Mixing corresponds to the degree of correlation between successive values in Markov Chain Monte Carlo. An optimally mixed Markov chain will yield successive values that are independent samples from the associated posterior distribution. In a Markov chain with poor mixing, successive values are not independent and many iterations may be required for an unknown parameter to leave a particular region of its associated marginal posterior distribution. Convergence describes whether the set of samples is collectively representative of the associated posterior distribution. Short Markov chains with poor mixing will be heavily influenced by the specified initial values and may only represent a small region of the posterior distribution. Common strategies to avoid poor convergence include burn-in (*i.e.* not using the results from a specified number of iterations at the outset of a Markov chain to describe the posterior distribution) and generating very long Markov chains (thousands or millions of iterations can sometimes be generated relatively quickly). Evaluation of mixing and convergence for specific examples is addressed in Sections 4.4.3, 5.3.3, and 6.3.2.

3.3 USE OF PROBABILISTIC MODELS TO COMPARE EXPERIMENTAL DESIGNS

Probabilistic models can be used as a framework to draw inferences about unknown model parameters from available data (*e.g.* using likelihood or Bayesian approaches) or to investigate the expected properties of hypothetical or future data given specified parameter values. The latter type of analysis, for example, can be used to explore properties of data associated with a particular experimental design (*e.g.* to determine whether a proposed experiment can yield accurate estimates of the parameters of interest) and to evaluate the distribution of various statistics (*e.g.* the sampling distribution of the mean or variance).

3.3.1 Monte Carlo

When the output of a model is a deterministic function of parameters (or a random variable with distribution parameters) that are themselves random variables, it is sometimes difficult to evaluate the marginal distribution of the output variable explicitly. Monte Carlo simulation is typically used when integration of the variability of all the random variables in the model is intractable. It is a process in which the random variables in a model are sequentially generated using their respective distributions to yield a single random model output, and this simulation procedure is often repeated many times. For example, in a hierarchical probabilistic model in which the distribution of the random variable X is a function of the parameter Y (which is itself a random variable described by a distribution) the marginal distribution of X may be of interest. This distribution can be obtained explicitly by integration across all values of Y (and their respective probabilities), but the integration may be difficult. Monte Carlo simulation repeatedly generates a random value of Y and a single random value of X conditional on the generated value of Y, and the set of generated X values represents a random sample from the marginal distribution of X.

Monte Carlo simulation can be used in conjunction with probabilistic models that describe the random errors in the process of obtaining enumeration data to simulate data or to investigate experimental design. In Section 6.5.2, for example, a simulated dataset is generated for each of several scenarios to evaluate the correlation between enumeration data and analytical recovery. Repeated Monte Carlo simulation of a particular experimental design or data analysis process can be used to compare alternative experimental designs or data analysis procedures. For example, Monte

Carlo simulation is used in Section 4.3 to investigate the design of recovery experiments that are used to estimate the mean and standard deviation of analytical recovery. It is also used in Section 6.2 to compare alternative concentration estimates that may be calculated from enumeration data.

3.3.2 Variance Decomposition

Variance decomposition is a tool that is used to evaluate the relative contribution of different sources of variability to a random variable (*e.g.* enumeration data). It is based on the expectation of random variables (*e.g.* E[X] = μ_X and Var[X] = E[($X - \mu_X)^2$] = σ_X^2). If X is a random variable that depends on Y and Y is also a random variable, then the overall mean of X can be calculated using Equation 3.11 and the overall variance of X can be calculated using Equation 3.12. The latter is known as the variance decomposition formula.

$$E(X) = E(E(X \mid Y)) \tag{3.11}$$

$$Var(X) = E(Var(X \mid Y)) + Var(E(X \mid Y)) \tag{3.12}$$

The hierarchical probabilistic models used to describe random errors in this thesis describe the data (X) as a function of model parameters (*e.g.* source concentration, analytical recovery of the sample) that are also (in some cases) random variables. Use of these equations enables evaluation of bias (*i.e.* determination of the difference between the expected value of a parameter estimate and the parameter's actual value) and evaluation of the relative contributions of different sources of error. As shown in Chapter 8, this knowledge can be used to design experiments that will enable collection of more precise data (by reducing the most important sources of variability) and to demonstrate more appropriate statistical analysis strategies (*e.g.* strategies to obtain unbiased parameter estimates and to correctly address measurement errors).

This type of analysis is the culmination of a detailed statistical analysis of a particular type of experiment. When research has expanded beyond case-specific data analysis approaches to development of more generally applicable models (that appropriately describe all sources of variability in the data) it is not only possible to conduct informed analysis of existing data, but also to design future experiments that will yield more informative data (or consume less resources).

Chapter 4

Analytical Recovery: Calibrating Counts to Actual Quantities of Analyte Particles

In any measurement process, it is important to consider how well measurements correspond to the actual values of the parameters they represent. In many cases, a conversion from raw measurements to unbiased estimates of the actual parameter value (*i.e.* calibration) is needed. This is especially true for methods in which measurements have different units than the parameter being measured (*e.g.* calibration of fluorescence measurements to concentrations of fluorescent particles). Calibration involves measurement of standards in which the parameter value (*e.g.* the quantity of analyte particles in the sample) is precisely known and development of a one-to-one function that enables estimation of the actual parameter value associated with a given measurement.

Microorganism and particle data obtained by enumeration often require calibration because the quantity of observed analyte particles is not necessarily equal to the actual quantity of analyte particles. For example, *Cryptosporidium* oocyst concentration estimates (in 'observed oocysts'/L) are usually systematically lower than the actual concentration (in oocysts/L) because the enumeration methods are known to include many losses. Quantification of analytical recovery, "the capacity of the analyst to successfully count each microorganism or particle of interest in a sample using a specific enumeration method" (Schmidt *et al.*, 2010a) is necessary to calibrate enumeration-based concentration estimates to actual concentrations.

Analytical recovery is also important for method development and validation, especially for methods that are known to be prone to substantial losses (*e.g.* USEPA, 2005b). Poor and variable analytical recovery contributes to poor sensitivity (*i.e.* increased abundance of false-negative samples) and greater measurement error in analyte particle counts or concentrations. Accordingly, it is desirable to develop analytical methods with better recovery (*i.e.* recovery that is closer to 100% and less variable among samples) and to ensure that laboratories using a standardized method have acceptable and consistent performance.

The objectives of this chapter are as follows:

- to summarize strategies that are used to quantify analytical recovery and factors that may affect it (Section 4.1),
- to develop probabilistic models that describe random errors in various types of recovery data (Section 4.2),

- to investigate experimental design factors that may enable analysts to obtain more precise recovery estimates and more accurate assessments of the variation in analytical recovery among replicates (Section 4.3), and
- to describe and compare several ways in which replicate recovery data can be statistically analyzed (Section 4.4) including point-estimation of parameters for distributions that describe variability in recovery and Bayesian analysis of the uncertainty in these parameters.

The chapter is concluded with a discussion of practical implications (Section 4.5). Case studies are woven throughout this chapter to exemplify various concepts and to illustrate use of the proposed quantitative approaches.

4.1 APPROACHES TO QUANTIFY ANALYTICAL RECOVERY

This section summarizes various considerations in the analysis of analytical recovery. In particular, it summarizes common approaches to prepare seed doses and introduces the concept of seeding error, it describes several types of experiments to investigate analytical errors in enumeration methods, it discusses factors that may affect the representativeness of recovery data, and it reviews common strategies to quantify recovery in environmental samples. Several case studies are provided to illustrate concepts and to provide examples of enumeration methods with which analytical recovery is an important consideration. This section focuses upon general approaches and associated modelling implications. It does not provide a comprehensive survey of details for various enumeration methods and analyte particles. Much of the discussion addresses the analytical recovery of methods that are used to enumerate the protozoan parasites *Cryptosporidium* and *Giardia* in water because these methods often have poor and highly variable analytical recovery. Nonetheless, similar considerations may also apply to the preparation of seeded standards and the analysis of analytical recovery for other types of microorganisms or particles and for other types of media.

4.1.1 Preparation of Seed Doses

The two general approaches to prepare seeded standards are to add a precisely known number of analyte particles to the sample or to inoculate the sample with an aliquot of known volume from a stock suspension of known concentration. Although the number of seeded particles is commonly assumed to be precisely known, it is often subject to random ‘seeding errors’.

Preparation of seeded standards in which the number of analyte particles is precisely known is necessarily dependent upon a seeding methodology in which the seed dose is precisely enumerated and added to the sample without losses. Preparation of seed doses with precisely enumerated numbers of microorganisms or microscopic particles is often dependent upon cell-sorting flow cytometry, although even this method is prone to some error (*e.g.* Reynolds *et al.*, 1999). Many commercially available seed doses such as EasySeed®, ColorSeed®, and BioBall® (BTF Precise Microbiology, Inc.; Pittsburgh, PA) report a mean and standard deviation for batches of seed doses: the actual mean dose may vary slightly from the nominal dose and there is variability among doses from a single batch. Quantitative analyses and models (including some discussed in this thesis) often regard the seed dose as precisely known (*e.g.* equal to the reported batch mean) because the distribution describing these seeding errors is generally unknown. Overlooked seeding error can result in some bias if the actual doses are systematically different from the assumed doses and will result in over-estimated recovery variability because some of the variability in replicate recovery estimates is due to seeding error rather than variable recovery. Variance decomposition is used to investigate the impacts of seeding error (without any distributional assumptions) upon recovery estimates in Sections 8.1.1 and 8.1.2.

An alternative method to prepare seeded standards is to seed them with an aliquot of stock suspension so that the concentration in the standard is approximately known. This approach is dependent upon having a stock of precisely known concentration and withdrawing a precisely known volume from the well-mixed stock. It is improbable, however, that the stock concentration is known exactly because there is always some uncertainty in a measured concentration (even if a rigorous method such as flow cytometry is used). Any error in the presumed stock concentration will result in biased recovery estimates, although the magnitude and direction of this bias would be unknown. Furthermore, the concentration may drift slightly as multiple doses are withdrawn from the stock (even if it is well-mixed) as demonstrated in Section 4.2.3. If the stock suspension is not well-mixed (*e.g.* if it is substantially clumped or settled), then the seed dose cannot be predicted with any accuracy. The probabilistic models presented herein assume that the seeding error in this approach is Poisson-distributed (*e.g.* Sections 4.2.3, 4.2.4).

Any losses between preparation of the seed doses and inoculation of the samples will result in under-estimated recovery and over-estimated variability of recovery. This additional type of seeding error has been evaluated in the preparation of feces samples that are seeded with *Schistosoma mansoni* eggs (Section 4.1.2; Franceschina *et al.*, in progress). To limit losses in the seeding methodology in Method 1623 (USEPA, 2005b), the seeding process is meticulously standardized. For example, tubes containing standardized seeding suspensions are vortexed before the seeding suspension is poured into the sample, then the tube is rinsed (with additional vortexing) and the rinsate is added to the seeded sample. This helps to ensure that (oo)cysts are not lost in the tube. When seed doses are withdrawn from a stock suspension by pipette, the pipette is pre-rinsed with a surfactant (to reduce attachment) and it is also repeatedly filled with the desired volume of stock suspension five times before drawing an aliquot with which to spike the sample.

4.1.2 Case Study: Errors in Seeding *Schistosoma mansoni* Eggs into Human Feces

Schistosoma mansoni is a human parasite that is diagnosed by detection of its eggs in feces. Seeded standards are used to evaluate the recovery or sensitivity of detection methods (*e.g.* Teixeira *et al.*, 2007; Sections 4.4.4 and 7.2). The eggs are sufficiently large (approximately 140 by 60 μm) that they can be aspirated individually from a microscope slide under microscopic observation and then seeded into the feces samples. Franceschina *et al.* (in progress) have conducted experiments to evaluate seeding errors when precisely enumerated doses of eggs were aspirated from a microscope slide and transferred to another microscope slide for re-counting. Average losses in the range of 10-20% (in doses ranging from 1 to 20 eggs) were observed. It was demonstrated, by enumerating the microscope slide from which the suspension had been aspirated, that not all of the eggs were aspirated from the slide and that there were also losses during the transfer (presumably eggs attached to, or retained in, the pipette tip). Accordingly, additional errors can be introduced to recovery or sensitivity data through losses in the seeding process. Possible reduction of these errors by rinsing the pipette tip and slide, then adding the rinsate to the seeded sample, has not presently been evaluated.

4.1.3 Types of Recovery Experiments

In general, analytical recovery is quantified by seeding samples with known quantities of analyte particles, processing the samples using typical methodology, and comparing the numbers of observed particles to the respective numbers of seeded particles. These experiments yield information about the analytical recovery of the method that can be expected when similar methodology is used to enumerate similar unseeded samples. Recovery experiments are also commonly completed with variations in water quality attributes, sample attributes, methodology, or analyst (or laboratory) to evaluate the impacts of various factors upon analytical recovery (as discussed in Section 4.1.5). Many of these experiments have been conducted in method development research (*e.g.* to identify which factors may compromise the recovery of a method, or to advocate use of particular methodology or equipment).

Method development (which may include minor modifications to enhance the quality control of an existing method or proposal of major changes in methodology) can be enhanced by having more information about the analytical errors in an enumeration method. For example, it may be useful to evaluate which components of a method are most susceptible to losses and to assess counting errors. Knowledge about the leading causes of analytical error enables targeted research and method development and yields results that are based on a scientific investigation of methodology rather than potentially circumstantial evidence of improved results. Detailed information about the analytical errors in a method can be obtained in two general ways: partial method recovery analysis or loss-tracking experiments.

In partial method recovery experiments, samples are seeded at a particular point in the sample processing methodology so that only the analytical errors associated with the subsequent methodology are measured. If samples seeded before a particular component of the methodology have substantially lower recovery than samples seeded afterwards, then that component of the method is responsible for the difference in analytical recovery. An experiment of this type was conducted by Nieminski *et al.* (1995) in which four seeded raw water samples were enumerated at each of three stages in the sample processing method using each of two alternative enumeration methods for *Cryptosporidium* and *Giardia*. In these experiments, it was determined that the sample collection and concentration step was responsible for most of the losses while the sample purification step and the membrane filtration and enumeration step were responsible for a smaller portion of the losses.

Loss-tracking experiments relate to the system model of the enumeration process that was described in Section 2.2 (*e.g.* Figure 2.2), in which each analyte particle has a source and a sink. In an ideal method, all counts result from actual analyte particles contained in the original sample (*i.e.* there is no contamination, growth, over-counting error, or false-positive counting) and all analyte particles in the sample are successfully enumerated (*i.e.* there are no losses, under-counting errors, or false-negatives). The losses associated with possible sinks in the method can be quantified by separately recovering and enumerating lost analyte particles (*e.g.* enumerating samples of discarded filtrate or supernatant or of rinse water). In some cases, it is possible to compare these losses to other measured sinks and to the final enumerated sample to determine the relative analytical error of particular components of the method. An example of a loss-tracking experiment concerning the enumeration of *Bacillus subtilis* spores and *Escherichia coli* using a method with serial dilutions and membrane filtration agar plates is provided in Section 4.1.4.

Compared to investigation of losses in an enumeration method, evaluating counting errors is simple because it is often possible for an analyst (or multiple analysts) to enumerate prepared samples several times (*e.g.* USEPA, 2005b). Such investigations help to determine how much of the variability in recovery data may arise from counting errors rather than poor sample processing methodology, and to compare the capability of various analysts. Unfortunately, it may not be possible to determine whether counting errors increase or decrease counts because it is often not possible to determine the correct number of enumerable particles (especially when identification errors are problematic).

4.1.4 Case Study: Losses of Culturable Microorganisms in Plating Methods

In plating methods (in which bacteria are enumerated by growing colonies in or on solidified nutrient media, or viruses are enumerated by developing plaques in solidified nutrient media containing large quantities of host cells), losses are generally not considered. It is known that the results are susceptible to considerable error because replicate analyses often yield highly variable results and there is a general belief among microbiologists that a difference below one order of magnitude is insignificant (*i.e.* such differences could arise from errors in the methodology alone). Furthermore, it is known that counts may be low because some of the target cells in a suspension may be viable but non-culturable and because some of the bacterial colonies may overlap or the viral plaques may be confluent. When using plating methods, analytical recovery should be investigated to determine if there are systematic losses that necessitate calibration of observed results to actual microbial concentrations and to evaluate the quality control of the methodology. Such assessments of recovery, however, are complicated by culturability: the ability to grow enumerable colonies or plaques from the target bacteria or viruses in a sample under specific laboratory conditions. While imperfect culturability is a loss that can be regarded as a component of analytical recovery, it may be non-random due to uncontrollable variations in the quality of the bacteriological media or the condition of the microorganisms in the sample.

An experiment was undertaken to investigate random errors in plate-count enumerations of *Bacillus subtilis* spores and *Escherichia coli* (Schmidt and Emelko, 2007). The experiment investigated dilution errors and losses of culturable bacteria during sample preparation. In plating methods, serial dilutions are commonly used to reduce microbial concentrations until an enumerable quantity of microorganisms can be plated. Errors associated with these dilutions can include inaccurate dilution factors (which will result in biased concentration estimates), unrepresentative sub-sampling (*i.e.* inadequate mixing of dilution tubes), and losses. In the membrane filtration procedure, possible losses of culturable microorganisms can include losses in the pipette tips, losses in the dilution tubes or filtration apparatus (*e.g.* droplets of water that are not filtered, or microorganisms that are attached to the dilution tubes or apparatus), bypass or breakthrough of the filter paper, and colony overlap on the plated filter.

Stock suspensions of *Bacillus subtilis* (ATCC 6633, Manassas, VA) spores and *Escherichia coli* (ATCC BAA-769, Manassas, VA) cells with concentrations of approximately 100 cfu[1]/mL were prepared and subsequently refrigerated to maintain relatively stable concentrations. During the stock preparation process, the *B. subtilis* spore stock was heat-shocked to eliminate vegetative cells. The experiment conducted with each microorganism consisted of two sets of five replicate dilution series, with the two sets completed one day (*B. subtilis*) or two days (*E. coli*) apart. In each case, preliminary results from the first set were used to modify the design of the second set (*i.e.* plates that were expected to yield no colonies were excluded to conserve time and resources).

For each dilution series, a 1 mL aliquot of well-mixed stock suspension was transferred into a dilution tube containing 9 mL of sterile 0.1x phosphate-buffered saline (PBS) using an Eppendorf pipette (Eppendorf AG, Hamburg, Germany) and a disposable 1 mL pipette tip (resulting in a 10^{-1} dilution). The pipette tip was then rinsed by repeated filling and discharging in a dilution tube of 0.1x PBS (the 10^{-1} pipette tip rinse) and subsequently discarded. Following vigorous vortexing, 1 mL of the 10^{-1} dilution was transferred to a 10^{-2} dilution tube (with similar rinsing of the pipette tip in some dilution series), and this process was repeated to prepare a 10^{-3} dilution. The entire contents of each of the prepared dilution tubes was filtered through sterile, 47 mm cellulose membrane filters with a 0.45 μm pore size (Pall Life Sciences, Port Washington, NY) using a 500 mL filtration unit (Nalgene model 300-4050, Rochester, NY). To reduce cross-contamination, the three dilutions were filtered in order of ascending concentration, and then the filtration unit was rinsed with sterile distilled water. The filters were placed in 50 mm Petri dishes containing nutrient agar. The nutrient agar used for the *B. subtilis* samples included 0.015 g/L Trypan Blue (as per Rice *et al.*, 1996). After filtering the three serial dilutions, the culturable microorganisms remaining in some dilution tubes (10^{-1}, and sometimes 10^{-2} and 10^{-3} also) were enumerated by filling the tubes with 9 mL of 0.1x PBS, vortexing, and plating the contents. The pipette tip and dilution tube rinse samples were plated using methodology similar to the main dilution series, after which the filtration unit was rinsed with sterile distilled water. A blank was filtered following this rinse after series C,E,H, and J. The collected filtrate and filtration unit rinse water was also filtered following the second set of *B. subtilis* samples and both sets of *E. coli* samples. All sample processing was conducted using aseptic technique to prevent contamination of the samples or agar plates. The Petri dishes were inverted and incubated for up to 48 hours at

[1] cfu = colony forming unit

temperatures of 30°C for *B. subtilis* and 37°C for *E. coli*. All plates were enumerated twice with the first count at approximately 24 hours and the second count at approximately 48 hours.

The results of the experiment indicated several possible errors that should be considered in plating experiments with dilution series: dilution errors and losses of culturable microorganisms. Despite careful volume control and ensuring that all dilution tubes were tightly capped during autoclaving, it was found that an average of 0.2-0.25 g of PBS had evaporated from each tube, which corresponds to a dilution factor of 9.75-9.80. The resulting bias, compounded over many sequential dilutions, can become quite high. Plate-count results for *B. subtilis* spores and *E. coli* are summarized in Tables 4.1 and 4.2. The two entries within each cell of the table represent the first and second count, and the sum column represents the total number of colony forming units recovered from each dilution series (all of which are presumed to have originated from the initial 1 mL sample used to prepare the series).

Culturable *B. subtilis* spores were recovered from many of the pipette tips used in the first dilution. Counts of these rinses were as high as 5 cfu, and averaged near 1.2% of the presumed number of colony forming units transferred (149 cfu). Fewer culturable *E. coli* were recovered from the pipette tips, with an average near 1.0% of the presumed number of colony forming units transferred (90 cfu). One of the *E. coli* pipette tip rinses (Series G) had an unusually high number of culturable

Table 4.1: *Bacillus subtilis* spore enumeration data (after Schmidt and Emelko, 2007)

Series	Dilution Series			Pipette Tip Rinses			Dilution Tube Rinses			Sum of Counts
	10^{-1}[a]	10^{-2}[b]	10^{-3}[c]	10^{-1}	10^{-2}	10^{-3}	10^{-1}	10^{-2}	10^{-3}	
A	120 121	4 5	2 3	4 5	0 0	0 0	5 5	0 0	0 0	135 139
B	143 144	16 16	0 0	0 0	0 0	0 0	5 6	0 0	0 0	164 166
C	150 150	13 14	1 1	2 2	0 0	0 0	4 4	0 0	0 0	170 171
D	109 112	14 15	0 1	0 3	0 0	0 0	4 4	1 1	0 0	128 136
E	140 139	10 10	0 0	1 2	0 0	0 0	1 2	0 0	0 0	152 153
F	113 109	8 9	0 0	1 2	-	-	2 3	-	-	124 123
G	147 145	11 12	0 0	3 4	-	-	8 8	-	-	169 169
H	140 129	14 15	0 0	0 0	-	-	4 4	-	-	158 148
I	123 128	13 15	2 2	4 3	-	-	2 2	-	-	144 150
J	121 122	15 15	1 1	0 1	-	-	3 3	-	-	140 142
Avg.	130.6 129.9	11.8 12.6	0.6 0.8	1.5 2.2	-	-	3.8 4.1	-	-	148.4 149.7

[a] 9 mL (0.9 mL of original sample) [b] 9 mL (0.09 mL of original sample) [c] 10 mL (0.01 mL of original sample)

Table 4.2: *Escherichia coli* enumeration data (after Schmidt and Emelko, 2007)

Series	Dilution Series			Pipette Tip Rinses			Dilution Tube Rinses			Sum of Counts
	10^{-1}[a]	10^{-2}[b]	10^{-3}[c]	10^{-1}	10^{-2}	10^{-3}	10^{-1}	10^{-2}	10^{-3}	
A	81	11	1	0	0	-	1	0	-	94
	79	11	1	0	0		1	0		92
B	81	8	1	0	0	-	0	0	-	90
	78	8	1	0	0		0	0		87
C	83	8	1	0	1	-	1	0	-	94
	84	8	1	0	0		0	1		94
D	70	8	0	0	0	-	2	0	-	80
	69	8	0	0	0		3	0		80
E	91	6	4	0	0	-	2	0	-	103
	84	6	4	0	0		2	2		98
F	70	12	1	1	-	-	1	-	-	85
	69	12	2	1			1			85
G	78	13	3	7*	-	-	1	-	-	102
	78	13	3	7*			1			102
H	84	8	1	0	-	-	3	-	-	96
	71	8	1	0			3			83
I	69	11	0	1	-	-	1	-	-	82
	73	11	0	1			1			86
J	84	8	2	0	-	-	2	-	-	96
	63	8	2	0			2			75
Avg.	79.1	9.3	1.4	0.9	-	-	1.4	-	-	92.2
	74.8	9.3	1.5	0.9			1.4			88.2
Avg. (No G)	79.2	8.9	1.2	0.2	-	-	1.4	-	-	91.1
	74.4	8.9	1.3	0.2			1.4			86.7

[a] 9 mL (0.9 mL of original sample) [b] 9 mL (0.09 mL of original sample) [c] 10 mL (0.01 mL of original sample)

* Presumed improper discharge of sample from pipette tip (*e.g.* large residual droplet)

microorganisms recovered from the pipette tip, possibly due to a large droplet of sample that was not discharged. Average results with and without Series G are included in Table 4.2 because this non-random error is only presumptive (*i.e.* the datum is not conclusively an outlier). The losses of *E. coli* in the pipette tips excluding Series G averaged 0.25% of the presumed number of colony forming units. Residual colony forming units recovered by rinsing the first dilution tube in each series were as high as 8 cfu for *B. subtilis* spores (with an average near 2.7%) and as high as 3 for *E. coli* (with an average near 1.6%). The combined losses associated with the first dilution averaged 3.9% of total colony forming units for *B. subtilis* and 2.5% for *E. coli* (1.9% with Series G excluded). In each of the experiments, one colony forming unit was observed among the four blanks (presumably indicating imperfect rinsing of the filtration unit between samples). Relatively few *B. subtilis* colony forming units (up to three) were recovered from the re-filtered filtrate, while the equivalent *E. coli* samples contained counts as high as 22 with evidence of contamination (*e.g.* atypical colonies).

The losses demonstrated in these two experiments indicate that raw concentration estimates based on plate counts are likely biased low. The counts that indicate losses, however, may be low because they only reflect lost microorganisms that were recovered by rinsing. Completing a similar experiment in which the quantities of microorganisms in the 1 mL samples is known (*e.g.* using precisely seeded standards or a more precisely enumerated stock) would enable estimation of the true overall losses. Additionally, the differences between the two counts of each plate indicated a need for better control of counting procedures or the time at which colonies are counted. Accordingly, analysis of losses and counting errors in plate-count methods should be a routine component of method or analyst validation and quality control, and it may be necessary to calibrate concentration estimates to actual concentrations through knowledge of analytical recovery.

4.1.5 Representativeness of Analytical Recovery Estimates

At present, the factors that affect analytical recovery in various enumeration methods and the extent to which these factors are case-specific are generally unknown. Accordingly, it is not known which factors need to be precisely controlled in order to obtain meaningful recovery estimates. Factors that could affect analytical recovery include water quality attributes (*e.g.* turbidity, temperature, pH), sample attributes (*e.g.* sample volume, analyte particle concentration), methodological attributes (*e.g.* apparatus, materials, reagents), and laboratory attributes (*e.g.* analyst capability). Crainiceanu *et al.* (2003) analyzed the nation-wide matrix spike (oo)cyst recovery data obtained from many water supplies and laboratories in accordance with the United States' Information Collection Rule. They used a Bayesian generalized linear mixed model to evaluate possible covariation between analytical recovery and various water quality and sample attributes (turbidity, temperature, pH, sample volume, and seed dose) and to account for laboratory and random effects. The sample volume was the only potential covariate that was found to have a substantial effect for either *Cryptosporidium* oocysts or *Giardia* cysts and laboratory effects (*i.e.* variation in recovery among the various laboratories used to obtain the ICR data) were only substantial for *Giardia*. This type of regression-based modelling approach is too complicated for many simple recovery data-sets, but is useful to evaluate numerous possible covariates simultaneously (when such data exist). Most of the research on factors that affect analytical recovery is based upon controlled experiments with batches of replicate recovery data. These have been used to evaluate many variations in methodology or sample attributes for (oo)cyst enumeration methods. Examples include Vesey *et al.* (1993), LeChevallier *et al.* (1995), Nieminski *et al.* (1995), Shepherd and Wyn-Jones (1995), Bukhari *et al.* (1998), Falk *et al.* (1998), McCuin *et al.* (2001), DiGiorgio *et al.* (2002), Feng *et al.* (2003), Massanet-Nicolau (2003), McCuin and Clancy (2003), and Ferguson *et al.* (2004). The abundance of literature addressing analytical recovery of various (oo)cyst enumeration methods associated with various types of water samples demonstrates that method development and investigation of factors that affect analytical recovery was (and possibly still is) a very active research area, although this has likely diminished since the development of Method 1623 (USEPA, 2005b).

In general, recovery data are only useful for inference about the recovery of indigenous (*i.e.* non-seeded) particles if the samples and methodology are representative of the enumeration of indigenous particles. Petterson *et al.*, 2007 summarized the (oo)cyst recovery data available for the twelve utilities involved in the MicroRisk Quantitative Microbial Risk Assessment project. Of these, five provided no recovery data at all, two provided recovery data for ultra-pure water only, three provided recovery data for environmental source water samples, one provided recovery for both types of samples, and one provided recovery data for ultra-pure water with added sediment. The extent to which recovery in synthetic water quality matrices is representative of environmental samples was raised as a source of uncertainty in the associated risk assessments. While recovery in synthetic water quality matrices has questionable relevance to recovery in environmental samples, it is useful for standardized quality control protocols because differences in performance among analysts or laboratories (or scatter in control charts over time as is addressed in Section 4.4.2) indicate differences in performance without the potentially confounding effects of variable water quality attributes. Accordingly, many of the validation criteria associated with Method 1623 (USEPA, 2005b), a standardized method to enumerate *Cryptosporidium* and *Giardia* in water, are based on seeded reagent water samples.

A final concern about the representativeness of recovery data is whether the analytical recovery of seeded and indigenous particles is equivalent. Possible problems include the acclimatization of seeded particles to the sample and physical differences between seeded and indigenous particles that may affect analytical recovery. Indigenous particles may have different recovery from seeded particles because they have had more opportunity to acclimatize to the chemistry of the sample and may be aggregated or attached to other particles. Additionally, differences in the surface characteristics of colloids are known to have a substantial impact on filtration and may also affect analytical recovery. Possible concerns about the representativeness of pre-stained ColorSeed® (oo)cysts (BTF Precise Microbiology, Inc.; Pittsburgh, PA) have been raised (Warnecke *et al.*, 2003). More recently, it has been demonstrated that matrix spike recovery can vary depending on the supplier of a particular isolate (Iowa strain) of *Cryptosporidium parvum* oocysts (Villegas, 2010). Investigation of the factors that affect analytical recovery and determination of strategies to obtain more appropriate recovery data are an ongoing research need.

4.1.6 Evaluating Analytical Recovery in Environmental Samples

Assessment of analytical recovery in environmental samples is essential because particle concentration estimates may be systematically different from actual concentrations and recovery data from other types of samples may not be sufficiently representative of the environmental samples to enable appropriate calibration. Quantifying analytical recovery in environmental samples is complicated, however, because there may be indigenous particles in the samples. Two general approaches have been adopted to account for background concentrations in (oo)cyst matrix spike samples: paired matrix spike data in which one sample is seeded to evaluate recovery while another is not to enumerate background oocysts (*e.g.* Teunis and Havelaar, 1999; USEPA, 2005b), and internal seeding using uniquely labelled (oo)cysts that can be enumerated separately from indigenous oocysts (*e.g.* ColorSeed®: Warnecke *et al.*, 2003).

4.1.7 Case Study: Matrix Spike Quality Control Requirements in Method 1623

The United States Environmental Protection Agency's Method 1623 (USEPA, 2005b) – enumeration of *Cryptosporidium* and *Giardia* in water by filtration, immunomagnetic separation, and immunofluorescence assay – is the currently accepted method for regulatory monitoring of (oo)cysts in water supplies. Method 1622 (USEPA, 2005a), an equivalent method in which only oocysts are enumerated, is not specifically addressed in this thesis because it is nearly identical to Method 1623 (although the matrix spike acceptance criteria are slightly different). In addition to method validation requirements based on assessment of recovery in reagent water (the initial and ongoing precision and recovery tests), the method's prescribed quality control program includes matrix spike (MS) samples and matrix spike duplicate (MSD) samples.

Matrix spike samples are required with a minimum frequency of 1 MS sample per 20 field samples (and at least one MS sample per week during which samples are enumerated) from each source analyzed. The matrix spike test requires paired samples (two concurrently collected samples or a split sample), in which one is seeded with oocysts and cysts and the other is not. The percent recovery is calculated using Equation 4.1, in which R is the percent recovery, N_{sp} and N_s are the numbers of oocysts or cysts observed in the spiked and unspiked samples respectively, and T is the number of oocysts or cysts added. Issues surrounding the accuracy and precision of this recovery estimate are addressed in Section 4.2.6. The acceptance criteria specify that the calculated matrix spike percent recovery must be between 13-111% for *Cryptosporidium* and 15-118% for *Giardia*.

$$R = \frac{N_{sp} - N_s}{T} \times 100\% \tag{4.1}$$

Matrix spike samples and matrix spike duplicates are recommended in Tier 1, and required in Tier 2, of the performance-based measurement system (which regulate method modifications for a single laboratory and nationwide approval respectively). The MSD sample is a field sample aliquot (identical to the MS sample) which is seeded, processed, and enumerated alongside the MS sample to evaluate precision of the matrix spike recovery (Clancy *et al.*, 2003). The relative percent difference (*RPD*) is calculated using Equation 4.2, in which N_{MS} and N_{MSD} are the number of oocysts or cysts counted in the MS and MSD samples respectively, and X_{MEAN} is the average of N_{MS} and N_{MSD}. The acceptance criteria specify that the calculated relative percent difference must be at most 61% for *Cryptosporidium* and 30% for *Giardia*. The count in the unspiked field sample is irrelevant in Equation 4.2 because the difference between the MS and MSD counts is of concern rather than the analytical recovery.

$$RPD = \frac{\left| N_{MS} - N_{MSD} \right|}{X_{MEAN}} \times 100\% \tag{4.2}$$

The seed doses used in matrix spike and reagent water samples (containing approximately 100-500 oocysts and 100-500 cysts) may be prepared by flow cytometry or from a manually enumerated stock (*e.g.* using hemocytometer chamber counting, well slide counting, or membrane filter counting). The acceptable relative standard deviation of the seed doses is up to 2.5% with flow cytometry, and 16% and 19% for *Cryptosporidium* and *Giardia* respectively with manual preparation. In some cases, it may be possible to substantially reduce the RPD simply by improving the precision of seed doses.

4.2 MODELLING RANDOM ERRORS IN RECOVERY DATA

The conventional recovery estimate is the fraction of seeded particles that are observed in a processed sample. In a sample that contains four particles, it may be possible to observe 0%, 25%, 50%, 75%, or 100% of the particles (or 125%, 150%, *etc*. if counting errors can lead to inflated counts). As described in Section 2.3, these recovery estimates have discrete values (because the number of observed particles is an integer) that are dependent upon the quantity of seeded particles (*e.g.* it is not possible to see 50% of one seeded particle). Analytical recovery is typically modelled as a probability or a rate to avoid this necessary dependence upon the quantity of particles in the sample because the quantity of particles present in unseeded samples (and many types of seeded samples) is not precisely known. Accordingly, analytical recovery should be regarded as a measure of the potential of the method to result in observation of any particles that may be present in a sample and not as the fraction of particles that are actually observed. The fraction of seeded particles that are observed is an imprecise estimate of analytical recovery.

Three random errors that affect recovery estimates are addressed herein: seeding error, analytical error, and non-constant analytical recovery. Seeding error represents variability between the actual number of particles seeded into a sample and the presumed seed dose. As described in Section 4.1.1, there may be variability in the numbers of particles in commercially available seed doses. There is also considerable variability when samples are seeded with an aliquot of precisely known volume that was withdrawn from a well-mixed stock of precisely known concentration (Section 4.2.3). These seeding errors contribute to the variability of recovery estimates and can result in biased recovery estimates if the average seed dose is different from the quantity that is assumed. For example, recovery estimates will be biased low if samples are systematically seeded with fewer particles than the assumed seed dose. Analytical error represents the difference between the number of observed particles and the number of particles actually present in a sample due to imperfect analytical recovery (and/or partial sample analysis). It is a random error because a method that is performed such that 50% of particles are likely to be observed will not necessarily result in observation of exactly 50% of the particles. Similarly, a small number of tosses of a fair coin will not always result in exactly 50% showing ‘heads’. Non-constant analytical recovery represents variation in recovery among samples (*e.g.* if 50% of the particles are likely to be observed in one sample and only 40% are likely to be observed in another due to an uncontrollable variation in methodology or sample characteristics).

This section presents six probabilistic models that use simple distributions to describe the random errors in various types of samples. These do not describe all possible random error models for recovery estimates; for example*,* only two alternatives are considered for seeding errors, only two non-constant analytical recovery distributions are considered, no model is developed for samples with precisely known numbers of seeded particles and recovery exceeding 100%, and possible covariation with other measurable values is not addressed. These models, however, are reasonably applicable in many situations and can be used to demonstrate the necessity of considering random errors in recovery estimates: both for enhanced experimental design and appropriate data analysis. Section 4.2.1 discusses several assumptions that are common to all of the models presented throughout this thesis. The following three sections introduce the three principal recovery models described in Schmidt *et al.*, 2010a. Section 4.2.2 addresses recovery data with precisely known numbers of seeded particles and beta-distributed non-constant analytical recovery (the 'beta-binomial recovery model'). Section 4.2.3 expands upon the beta-binomial recovery model by including Poisson-distributed seeding error when samples are seeded using an aliquot from a stock suspension (the 'beta-Poisson recovery model'). Section 4.2.4 provides an alternative to the beta-Poisson model in which non-constant analytical recovery is described by a gamma distribution that allows recovery values exceeding 100% (the 'negative binomial recovery model'). Section 4.2.5 presents modifications of the three preceding models to account for partial sample analysis (*e.g.* when a sample is subdivided so that only a portion is enumerated). Two models are developed to describe commonly used methods to quantify analytical recovery of *Cryptosporidium* oocysts and *Giardia* cysts in environmental samples (which may contain indigenous (oo)cysts). The first (Section 4.2.6) addresses random errors in paired matrix spike samples collected in accordance with the Method 1623 validation criteria (USEPA, 2005b; Section 4.1.7). The 'internal seed recovery model' (Section 4.2.7) addresses random errors when recovery is evaluated in environmental samples using ColorSeed® (oo)cysts, which are distinguishable from the indigenous (oo)cysts that may also be present in the sample. Table 4.3 provides a basic comparison of many of these recovery models. To conclude this section on models describing random errors in analytical recovery data, Section 4.2.8 provides a critical review of other probabilistic models that have been used to describe random errors in the enumeration process and contrasts them with the models presented herein.

4.2.1 Common Assumptions in Probabilistic Modelling of Analytical Errors

Several assumptions are common to all of the models that are presented throughout this thesis. It is assumed (1) that the analytical recovery of seeded particles is representative of the recovery of indigenous particles, (2) that all particles in the sample are equally likely to be observed, (3) that analytical recovery is independent of the number of seeded particles, (4) that the entirety of each sample is enumerated (unless partial sample analysis is specifically addressed in the model), and (5) that the method is in statistical control (*i.e.* there are no non-random errors). These assumptions are discussed below.

In general it is assumed that the analytical recovery of seeded particles is representative of the recovery of indigenous particles. The strategies to evaluate analytical recovery and the associated statistical analyses that are presented in this chapter quantify the analytical recovery of seeded particles, and this information is used in the subsequent chapters to make inferences about the recovery of indigenous particles (*e.g.* to calibrate concentration estimates to actual concentrations in Chapters 5 and 6 and to quantify sensitivity of detection methods in Chapter 7). Use of seeded particles is necessary because directly quantifying the analytical recovery of indigenous particles is generally infeasible: the number or concentration of indigenous particles in the sample would have to be precisely known *a priori* to quantify analytical recovery of indigenous particles. At present, the properties of analyte particles that affect analytical recovery are not well-known, so it is not possible

Table 4.3: Probabilistic models for random errors in recovery data

Model	**Beta-binomial** (Section 4.2.2)	**Beta-Poisson** (Section 4.2.3)	**Negative Binomial** (Section 4.2.4)	**Internal Seed** (Section 4.2.7)
Application	Samples seeded with known numbers of analyte particles, indigenous particles absent, recovery <100%	Samples seeded with known concentrations of particles, indigenous particles absent, recovery <100%	Samples seeded with known concentrations of particles, indigenous particles absent, recovery can be >100%	Samples seeded with known numbers of analyte particles, seeded and indigenous particles counted separately, recovery <100%
Seeding Error	–	Poisson	Poisson	–
Analytical Error	Binomial	Binomial		Binomial
Non-constant Analytical Recovery	Beta	Beta	Gamma	Beta (optional)
Counting Errors	Random counting errors included if counts never exceed seed doses	Random counting errors included if counts never exceed seed doses	Random counting errors included	Random counting errors included if counts never exceed seed doses
Partial Sample Analysis (Section 4.2.5)	Incorporated into binomial distribution if sample is well-mixed	Incorporated into binomial distribution if sample is well-mixed	Incorporated into Poisson distribution if sample is well-mixed	Incorporated into binomial distribution if sample is well-mixed

The model developed for paired matrix spike samples (Section 4.2.6) is not shown in this table. It is based on the beta-Poisson model (or beta-binomial if the number of seeded (oo)cysts is precisely known). It is assumed in this model that there is no variability in recovery within pairs, but that there is beta-distributed non-constant analytical recovery among pairs.

to address the potential differences in recovery between seeded and indigenous particles. Concerns about the representativeness of seeded particle recovery estimates are also discussed in Section 4.1.5.

The models presented in this thesis assume that all analyte particles in a sample have equal analytical recovery (*e.g.* an equal probability of being observed or rate of observations per particle). Different types of analyte particles that are enumerated concurrently may, however, have different analytical recovery. In Method 1623 (USEPA, 2005b), *Cryptosporidium* oocysts and *Giardia* cysts are enumerated concurrently and their analytical recoveries are generally unequal; this demonstrates that losses and/or counting errors can be affected by properties of the analyte particles. This thesis does not address possible covariation of analytical recovery for the various types of analyte particles in multiplex assays (*i.e.* samples in which several analytes are enumerated concurrently). The assumption that all analyte particles in a sample have equal analytical recovery implies that the particles are uniform. A mixed population of analyte particles (*e.g.* samples containing multiple genotypes of *Cryptosporidium* oocysts) that is enumerated as a single type of particles can be problematic if the various sub-populations have different analytical recoveries. If some sub-populations of analyte particles in a mixed population have higher analytical recovery than others, then the average recovery of the mixed population depends on the relative abundance of the various sub-populations, and the estimated analytical recovery is only representative of other samples with an identical mixed population. At this time, particle characteristics that affect analytical recovery are not well-known, and it is assumed that recovery estimates are representative of the recovery of indigenous particles (possibly including mixed populations) in similar samples that are processed using the same methodology.

Throughout this thesis, it is assumed that analytical recovery is independent of the number, or concentration, of analyte particles in the sample (*i.e.* that the mean or variance of analytical recovery or the non-constant analytical recovery distribution is independent of seed dose and concentration). If this is not true, then recovery estimates are only representative of the analytical recovery in samples with similar concentrations of analyte particles (Schmidt *et al.*, 2010a). The more important practical problem, if analytical recovery depends on the quantity of analyte particles in the sample, is that calibrating enumeration-based concentration estimates to actual concentrations (using information about analytical recovery) would be very complicated. In Chapter 5, the concentration of analyte particles in a source is inferred from the number of observed particles in the sample using information about the possible values of analytical recovery. If analytical recovery depended on the quantity of particles in the sample, then it would be necessary to simultaneously estimate the unknown quantity of particles that were collected in the sample and choose which available recovery data would be appropriate for that quantity of particles. Such a scenario would require much more recovery data, and a much more complicated statistical approach, than what is proposed herein. Moreover, matrix spike recovery estimates would not be representative of the analytical recovery of indigenous particles because the recovery in the seeded and unseeded samples would be systematically different, and internal seed recovery estimates would not be representative of the analytical recovery in unseeded samples. Experiments to investigate possible dependence between analytical recovery and the quantity of analyte particles in the samples are necessary to evaluate the validity of the independence assumption used herein and the potential value of developing more complicated models and statistical analysis approaches.

In general, all models in this thesis assume that the entirety of each sample is enumerated. Examples in which the entire sample is not enumerated include partial pellet analysis in the ICR *Cryptosporidium* enumeration method (*e.g.* Young and Komisar, 1999) and Method 1623 (USEPA, 2005b) and field-of-view microscope slide enumerations for virus-size microspheres (Chae *et al.*, 2008). Discussion of how to modify the recovery models to include partial sample analysis is provided in Section 4.2.5. These models assume that the fraction of sample that is enumerated is precisely known and that the sample is well-mixed so that the subsample is representative of the entire sample. Non-random subsamples (*e.g.* from poorly resuspended pellets or from microscope slides upon which the particles are not randomly distributed) cannot be reliably used for inference because excessively variable results may arise from non-representative subsamples.

The models presented in this thesis also assume that the enumeration methods are in statistical control (*i.e.* that all errors are random and can be described by distributions as discussed in Section 3.1.1). Non-random errors including substantial analyst mistakes and predictable differences in recovery due to changes in particle characteristics, water quality attributes, sample attributes, or methodology are not addressed in these models (except in individual sample-specific recovery estimates obtained by internal seeding).

4.2.2 Probability of Recovery with a Precisely Known Number of Seeded Particles

This model, referred to as the beta-binomial recovery model throughout this thesis, addresses recovery in replicate samples that are seeded with precisely known numbers of particles. When the number of particles seeded into each sample is precisely known, there is no seeding error in the associated recovery estimates; only analytical error and non-constant analytical recovery need to be considered. In this model, analytical recovery is regarded as the probability that each particle in a sample that is processed by a particular method will be observed. The fraction of seeded particles that are observed is an estimate of this probability. This model is modified to account for partial sample analysis in Section 4.2.5.

If it is assumed that all of the particles in a sample have an equal probability of being observed and that the observation of each particle is independent of the loss or observation of all other particles, then the particles are observed or lost according to a Bernoulli process and the number of observed particles can be modelled by a binomial distribution that represents analytical error (Nahrstedt and Gimbel, 1996; Schmidt *et al.*, 2010a). This model technically assumes that all particles present in a sample are either observed or lost and that counting errors are limited to losses (*i.e.* there are no false-positive identifications or over-counting errors). Counts that exceed their respective seed doses are incompatible with this model. If it is assumed, however, that these counting errors are random and cannot inflate counts above the number of seeded particles, then they can be regarded as a component of analytical recovery and this model can still be used. Binomial models have been used extensively to determine the number of particles in a sample that successfully pass through some type of process:

examples include recovery in enumeration methods (Nahrstedt and Gimbel, 1996; Parkhurst and Stern, 1998; Teunis *et al.*, 1999; Teunis and Havelaar, 1999; Petterson *et al.*, 2007; Schmidt *et al.*, 2010a; Emelko *et al.*, 2010a), sub-sampling error (Parkhurst and Stern, 1998), and passage of microorganisms through treatment systems (Teunis *et al.*, 1999; Teunis and Havelaar, 1999).

It has commonly been assumed that the probability of recovery can vary randomly among samples and that this variation (*i.e.* non-constant analytical recovery) can be described by a beta distribution (*e.g.* Nahrstedt and Gimbel, 1996; Teunis *et al.*, 1997; Teunis *et al.*, 1999; Teunis and Havelaar, 1999; Crainiceanu *et al.*, 2003; Pouillot *et al.*, 2004; USEPA, 2005c; Signor and Ashbolt, 2006; Petterson *et al.*, 2007; Jaidi *et al.*, 2009; Cummins *et al.*, 2010; Schmidt *et al.*, 2010a; Emelko *et al.*, 2010a). The (standard) beta distribution is commonly used because it is bounded on the interval [0,1], it is practically unimodal (*i.e.* only unimodal beta distributions are of practical relevance), and it is the conjugate prior of the binomial distribution as described in Section 3.2.2 (Schmidt *et al.*, 2010a). Other distributions confined within this interval may provide a superior fit, but are not considered herein. Distributions that enable analytical recovery to exceed 100% (in which case recovery is no longer a probability) are incompatible with the binomial distribution used in this model.

The two distributions that are used in the beta-binomial recovery model yield the joint distribution shown in Equation 4.3. In this equation, *x* is the number of observed particles, *p* is the analytical recovery, *n* is the precisely known number of seeded particles, and *a,b* are shape parameters of the beta distribution for non-constant analytical recovery. The marginal beta-binomial distribution for the number of observed particles (*i.e.* upon integrating the nuisance parameter *p* out of the model) is described by Equation 4.4. The beta-binomial recovery model is implemented in recovery experiment design in Section 4.3 and in recovery experiment data analysis methods in Section 4.4. It is also used in the analysis of concentration data in Section 5.3.4 and in evaluation of the sensitivity of an enumeration method in Section 7.2.

$$f(x,p \mid n,a,b) = \left[\frac{n!}{x!(n-x)!} p^{x}(1-p)^{n-x}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}\right] \tag{4.3}$$

$$f(x \mid n,a,b) = \frac{n!\Gamma(a+b)\Gamma(x+a)\Gamma(n-x+b)}{x!(n-x)!\Gamma(a)\Gamma(b)\Gamma(n+a+b)} \tag{4.4}$$

Like many models, the assumptions described above can be relatively restrictive and it is important to understand the implications of using the model under non-ideal conditions. If the number of seeded particles is not precisely known, then two forms of bias can result. First, the recovery estimates (and the mean analytical recovery) can be biased if the actual numbers of seeded particles are systematically different from the presumed numbers of seeded particles. Even if prepared seed doses contain precisely known numbers of particles, losses during the seeding process (*e.g.* as demonstrated in Section 4.1.2) can result in bias because the excess losses are falsely attributed to analytical recovery. The second possible form of bias is over-estimated recovery variability (as demonstrated in Sections 4.3 and 8.1.1) because any excess variability in recovery measurements due to seeding error will be falsely attributed to non-constant analytical recovery. Monte Carlo simulations such as those presented in Section 4.3 can be used to evaluate the suitability of this model under various seeding error scenarios. In some cases, relatively low levels of seeding error may not have a substantial impact upon the recovery data and can be ignored. It is not strictly necessary for each sample described by this model to contain the same number of seeded particles (unless it has been demonstrated that recovery and the quantity of seeded particles are dependent). The model only requires that the number of seeded particles in each sample is precisely known. As described above, this model cannot be used if the number of particles observed in any sample exceeds the associated number of seeded particles. In addition to partial sample analysis (as described in Section 4.2.5), such an error may arise from unaddressed seeding error, contamination, growth of microorganisms, or counting errors that inflate the number of observed particles (*e.g.* over-counting and false-positive observations). Seeding error is addressed in Section 4.2.3 and a model that accounts for random counting errors that may sometimes cause analytical recovery to exceed 100% is discussed in Section 4.2.4. Contamination and growth are likely non-random errors and (like non-random counting errors) cannot be modelled.

4.2.3 Probability of Recovery with Particles Seeded from a Stock Suspension

This model, referred to as the beta-Poisson recovery model throughout this thesis, expands upon the previous model by incorporating Poisson-distributed seeding error. This particular form of seeding error is relevant because seeding samples with aliquots of known volume from a suspension of known concentration will result in an error that is often approximately Poisson-distributed (given some basic assumptions). In this model, counts that exceed their associated presumed seed doses are attributed to seeding error. Expansion of this model to address partial sample analysis is discussed in Section 4.2.5.

If a stock suspension is well-mixed (such that the suspended particles are randomly distributed throughout the stock), then the probability that each particle is contained in a withdrawn aliquot is equal to the fraction of stock withdrawn, and the number of particles in the aliquot is binomially distributed. For example, if a well-mixed 10-litre stock contains 1000 particles and a 1-litre aliquot is withdrawn, then each of the 1000 particles has a 10% chance of being in the withdrawn aliquot and the number of particles in the aliquot is binomially distributed. This binomial distribution can be very accurately approximated by a Poisson distribution (Section 3.1.2) if the total number of particles in the stock is high and the fraction of stock that is withdrawn is low. A series of withdrawn aliquots will be Poisson-distributed (according to their respective volumes) if the stock is continuously well-mixed (*i.e.* not clumped or settled) and the aliquots are independent (Schmidt *et al.*, 2010a). The aliquots will be approximately independent if the stock concentration remains constant and the total withdrawn stock volume is not a large fraction of the initial stock volume. As the fraction of stock that has been withdrawn becomes large, the samples become increasingly dependent because the concentration in the remaining stock is a function of the quantity of particles that have been withdrawn. Figure 4.1 illustrates the drift in concentration that will occur as successive samples are withdrawn from a source that is mixed after each withdrawal. Repeated aliquots will be

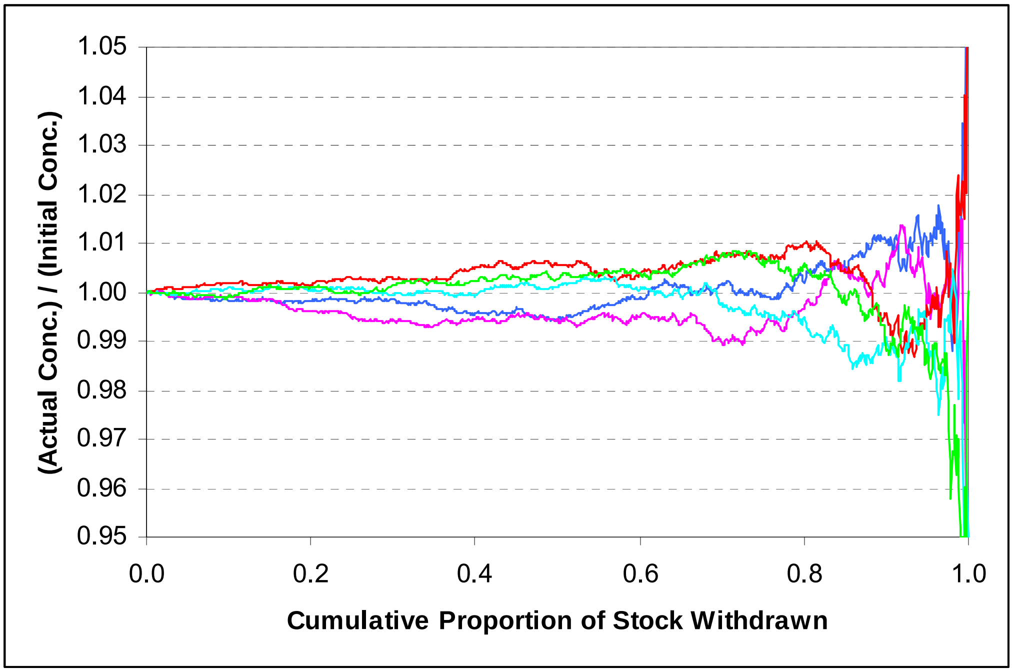


Figure 4.1: Stock concentration drift over successive withdrawals

Five example Markov Chains illustrating random stock concentration drift are shown. Each simulation is based on a stock containing 10^5 particles, with a total of 1000 aliquots withdrawn (the expected number of particles per aliquot is 100). A binomial model is used to simulate the number of particles in each withdrawn aliquot (with Poisson approximation where the number of particles is greater than 1000 and the fraction of stock being withdrawn is less than 5%). Mixing is assumed after each withdrawal so that the concentration becomes the remaining number of particles divided by the remaining volume.

non-representative and also particularly dependent if the stock is not well-mixed (*e.g.* an aliquot that non-randomly contains zero particles will increase the stock concentration by decreasing the volume in which the particles are suspended).

When Poisson-distributed seeding error can be assumed and is incorporated into the beta-binomial recovery model described in the preceding section, the joint distribution for the model is described by Equation 4.5. In this model, λ is the expected number of seeded particles (*i.e.* the product of the stock concentration and the volume seeded into the sample). This model can be simplified by removing the nuisance parameter n through summation to yield a Poisson distribution (Parkhurst and Stern, 1998) that accounts for seeding and analytical error as shown in Equation 4.6 (Schmidt *et al.*, 2010a). In this model, individual recovery estimates (x/λ) can exceed 100% due to seeding errors (*i.e.* if the count x is greater than the expected seed dose λ because the actual seed dose n is randomly greater than λ). This model is implemented in recovery experiment design in Section 4.3 and in recovery experiment data analysis methods in Section 4.4.1. A similar model is used to describe random errors in enumeration data (in unseeded samples) in Section 5.1.1.

$$f(x,n,p \mid \lambda,a,b) = \left[\frac{e^{-\lambda}\lambda^{n}}{n!}\right]\left[\frac{n!}{x!(n-x)!}p^{x}(1-p)^{n-x}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}\right] \tag{4.5}$$

$$f(x,p \mid \lambda,a,b) = \left[\frac{e^{-\lambda p}(\lambda p)^{x}}{x!}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}\right] \tag{4.6}$$

Seeding errors that are not exactly consistent with the Poisson model used herein can lead to bias. If the stock concentration or the volume of stock seeded into the sample are not precisely known, then the expected seed dose λ can be incorrect and the recovery estimates can be biased. Additionally, if the number of particles in the stock suspension is low or the fraction of stock withdrawn is large, then the seeding error is binomial rather than Poisson (which results in less variable seeding error than the model). Conversely, if the particles in the stock suspension are clumped (*i.e.* not randomly distributed), then the seeding error will be more variable than the associated Poisson model. If the assumed variability of the seeding error is inaccurate, then the estimated variability of non-constant analytical recovery will be biased.

4.2.4 Rate of Recovery with Particles Seeded from a Stock Suspension

When analytical recovery is regarded as a rate rather than a probability, and it is allowed to exceed 100% due to counting errors, the binomial model for analytical errors cannot be used. At present, no model with precisely known numbers of seeded particles and recovery possibly exceeding 100% has been developed. When the number of seeded particles is Poisson-distributed, however, the rate of observed particles per unit of stock volume is the product of stock concentration and recovery (*i.e.* the rate of particles per unit of stock volume multiplied by the rate of observations per particle). This model is analogous to the model presented in Equation 4.6, except that the rate of recovery is not a probability and is modelled by a gamma distribution. It is based upon a similar model (Margolin *et al.*, 1981) in which it was assumed that variability in the Poisson distribution mean among *Salmonella* plate counts, due to random errors in the enumeration procedure, could be described by a gamma distribution. The model presented herein is different because it addresses analytical recovery in seeded samples rather than sample-to-sample variation in the Poisson mean of indigenous microorganism counts, and also because the Poisson mean is separated into its various components (*i.e.* the expected number of particles in the sample and analytical recovery). It is convenient to model non-constant analytical recovery with a gamma distribution because it is confined to values greater than zero and allows analytical recovery to exceed 100%, because it is unimodal and has relatively flexible shape, and because it is the conjugate prior of the Poisson distribution as described in Section 3.2.2 (Emelko *et al.*, 2010a). Other distributions may provide a superior fit, but are not considered herein. Modification of the model to include partial sample analysis is discussed in Section 4.2.5.

The joint distribution resulting from the Poisson distribution (which represents both seeding and analytical errors) and gamma-distributed non-constant analytical recovery (with scale and shape parameters α,β) is described by Equation 4.7. The marginal distribution for the number of observed particles (*i.e.* upon integrating the nuisance parameter p out of the model), is described by Equation

4.8 (Schmidt *et al.*, 2010b). This model is referred to as the negative binomial recovery model throughout this thesis because Equation 4.8 is a negative binomial distribution (Section 3.1.2). This model is implemented in recovery experiment data analysis methods in Section 4.4.1, and a similar model is used to describe random errors in enumeration data (in unseeded samples) in Section 5.1.2.

$$f(x,p\mid\lambda,\alpha,\beta)=\left[\frac{e^{-\lambda p}(\lambda p)^x}{x!}\right]\left[\frac{1}{\beta^\alpha\Gamma(\alpha)}p^{\alpha-1}e^{-p/\beta}\right] \qquad (4.7)$$

$$f(x\mid\lambda,\alpha,\beta)=\frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)}\left(\frac{\lambda\beta}{\lambda\beta+1}\right)^{\alpha}\left(\frac{1}{\lambda\beta+1}\right)^{x} \qquad (4.8)$$

4.2.5 Recovery Estimates with Partial Sample Analysis

In some methods, it is common to process a sample to a certain point in the methodology and then subdivide the sample for further processing. For example, if a pellet of more than 0.5 mL is obtained in the centrifuged filter eluate in Method 1623 (USEPA, 2005b), then the pellet is resuspended and a volume equivalent to not more than 0.5 mL of the pellet is subjected to immunomagnetic separation. Partial sample analysis was also common in the ICR *Cryptosporidium* enumeration method, and was incorporated into several types of models for enumeration and recovery data (*e.g.* Parkhurst and Stern, 1998; Young and Komisar, 1999; Crainiceanu *et al.*, 2003).

As described in Section 4.2.3, the probability that each particle present in a well-mixed suspension is contained in a subsample is equal to the fraction of the total volume comprising the subsample (θ). Additionally, each particle in the subsample is assumed to have an equal probability of being observed (p), so that each particle in the original sample will have probability θp of being observed in the subsample (Parkhurst and Stern, 1998). This probabilistic model is described by nested binomial distributions (as shown in Equation 4.9), which simplify to a single binomial distribution (as proven in Appendix A). A similar model is implicitly assumed when the binomial distribution is used for other analytical errors because each step in the enumeration process can be modelled by a binomial distribution and the overall result is also a binomial distribution (with an overall probability of recovery equal to the product of the recoveries of each of the steps).

$$\sum_{m=x}^{n}\left[\binom{n}{m}\theta^m(1-\theta)^{n-m}\right]\left[\binom{m}{x}p^x(1-p)^{m-x}\right]=\binom{n}{x}(\theta p)^x(1-\theta p)^{n-x} \qquad (4.9)$$

As a result, Equations 4.10 and 4.11 are obtained when the models for known and Poisson-distributed numbers of seeded particles (Sections 4.2.2 and 4.2.3) are modified to include partial sample analysis (Schmidt *et al.*, 2010b). When partial sample analysis is incorporated into the negative binomial model presented in Section 4.2.4, the rate of observations per particle present in the sample becomes θp, which results in the model described by Equation 4.12 (Schmidt *et al.*, 2010b).

$$f(x,p\mid n,\theta,a,b)=\left[\frac{n!}{x!(n-x)!}(\theta p)^x(1-\theta p)^{n-x}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}\right] \qquad (4.10)$$

$$f(x, p \mid \lambda, \theta, a, b) = \left[\frac{e^{-\lambda\theta p}(\lambda\theta p)^x}{x!}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}\right] \tag{4.11}$$

$$f(x, p \mid \lambda, \theta, \alpha, \beta) = \left[\frac{e^{-\lambda\theta p}(\lambda\theta p)^x}{x!}\right]\left[\frac{1}{\beta^{\alpha}\Gamma(\alpha)} p^{\alpha-1} e^{-p/\beta}\right] \tag{4.12}$$

In addition to assuming that the particles are randomly distributed throughout the sample when it is subdivided, these models assume that analytical recovery is independent of the fraction of sample that is analyzed. The assumption that particles are randomly distributed is reasonably appropriate if the sample is well-homogenized (*e.g.* during resuspension of the pellet). Clumping or other non-random distributions of particles throughout the sample when it is subdivided will result in over-dispersed sub-sampling error as discussed by Young and Komisar (1999). Discussion of other probabilistic models that have been used to analyze recovery data, including several that address partial sample analysis, is included in Section 4.2.8.

4.2.6 Matrix Spike Recovery Estimates in Method 1623

In order to obtain recovery estimates in environmentally relevant samples, natural samples can be seeded with known quantities of particles and enumerated using typical methodology. These matrix spike recovery estimates are complicated by the potential presence of indigenous particles in the samples. The conventional recovery estimate, which is the number of observed particles divided by the number of seeded particles, would be biased high if any of the observed particles were indigenous. Accordingly, the matrix spike recovery procedures in Method 1623 (USEPA, 2005b) correct matrix spike recovery estimates for the presence of indigenous (oo)cysts as shown in Equation 4.1. There are a number of possible random errors that must be considered, however, in analyzing matrix spike recovery data. Like the recovery models presented above, these matrix spike recovery estimates may include seeding and analytical errors, non-constant analytical recovery, and errors due to partial sample analysis. In addition to these errors, however, there is also error in whether or not the number of indigenous (oo)cysts observed in the unspiked sample is equal to the number of observed (oo)cysts in the spiked sample that are indigenous. This section does not present a comprehensive model of all possible errors in matrix spike recovery data (*e.g.* variability in recovery among pairs of samples, partial sample analysis, and recoveries exceeding 100% are not considered herein); rather, it presents a model to illustrate the possible errors that make matrix spike recovery data more complicated than the recovery data associated with the preceding models.

As discussed in Chapter 5, unequal counts of indigenous (oo)cysts can arise from different concentrations or sample volumes, or from random measurement errors (*e.g.* random sampling error, analytical error, or non-constant analytical recovery). A probabilistic model for these errors is presented in Figure 4.2. In this model, each sample (possibly including a matrix spike duplicate that is not shown) has a potentially sample-specific concentration. The samples, however, are all presumed to be representative of the same indigenous (oo)cyst concentration (*i.e.* $c_1 = c_2$) because they are

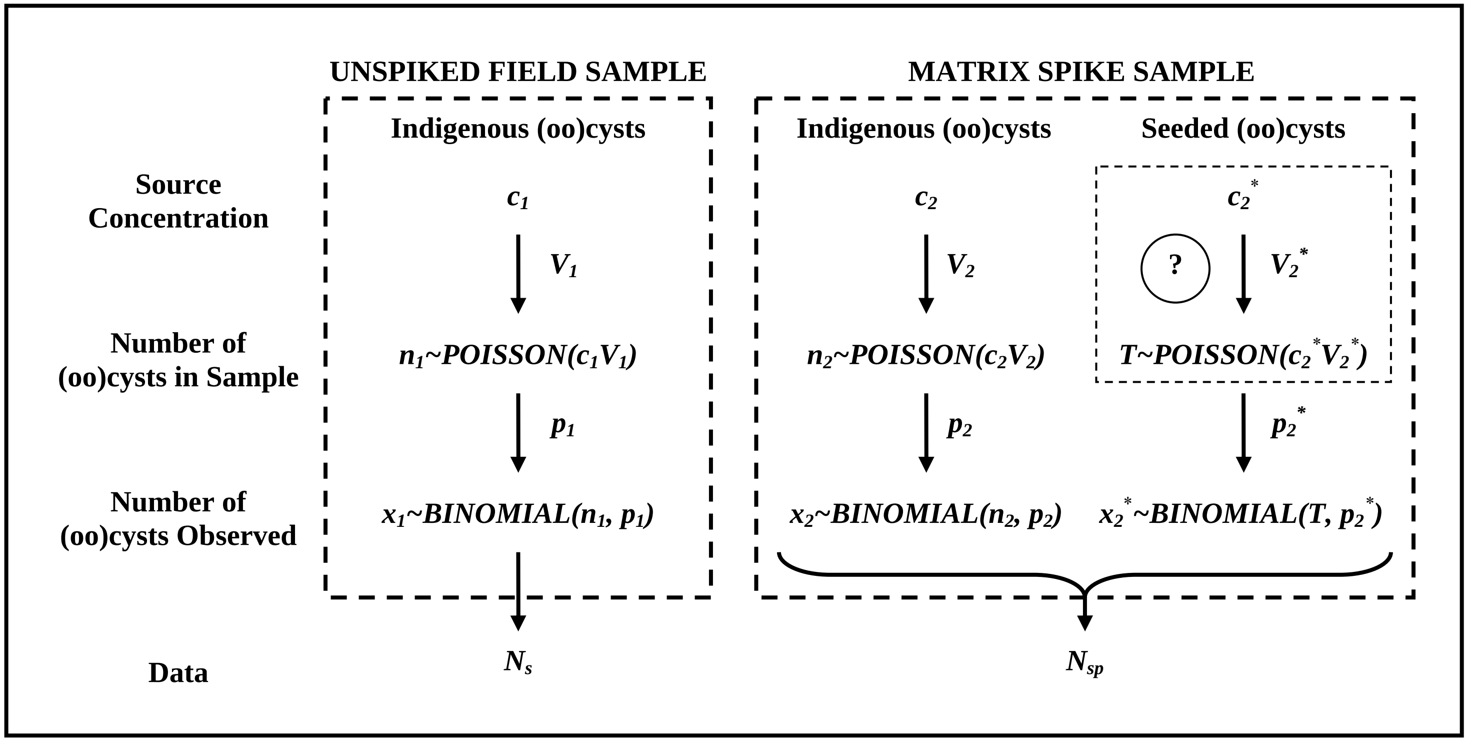


Figure 4.2: Probabilistic model for Method 1622/1623 matrix spike seeding

The model shows all the random errors that may contribute to variability in paired matrix spike counts. The number of seeded (oo)cysts (*T*) may be precisely specified by flow cytometry, or it may be Poisson-distributed as shown.

assumed to be replicates (*e.g.* a split sample or concurrently collected samples). Assuming that the indigenous (oo)cysts are randomly distributed (*i.e.* not clumped) and that the concentration is relatively homogeneous, the number of indigenous (oo)cysts in each sample can be modelled by a Poisson distribution in which the mean is equal to the product of concentration and sample volume. Samples with different volumes (*i.e.* $V_1 \neq V_2$) will lead to bias (when the count of the matrix spike sample is reduced by the count of the unspiked sample) because the expected number of indigenous (oo)cysts in each sample is proportional to its sample volume. Even if the samples are replicates with equal volume, however, random sampling error is unavoidable and it is improbable that the samples will contain equal numbers of indigenous (oo)cysts. Assuming that recovery is modelled as a probability, the number of observed indigenous (oo)cysts in each sample is binomially distributed as a function of the sample-specific number of indigenous (oo)cysts and probability of recovery. According to this model, the numbers of observed indigenous (oo)cysts can be unequal even if the two samples contain equal numbers of indigenous (oo)cysts (*i.e.* $n_1 = n_2$) and have equal probabilities of recovery (*i.e.* $p_1 = p_2$). Many modellers, however, have assumed that analytical recovery can vary randomly among replicate samples (*e.g.* according to a beta distribution), and this further complicates analysis of matrix spike recovery data. Recovery may also vary among matrix spike pairs because of changes in water quality.

In summary, biased matrix spike recovery estimates will result if the unspiked field sample and the matrix spike sample represent unequal concentrations, have unequal sample volumes, or have unequal mean analytical recovery. Additionally, there are many unavoidable random errors in the matrix spike recovery assessment procedure. Therefore, the number of indigenous (oo)cysts in the unspiked field sample is unlikely to be equal to the number of observed (oo)cysts in the matrix spike sample that are indigenous. The recovery estimate presented in Equation 4.1, however, is unbiased if (1) the samples are representative of the same source concentration (*i.e.* they are replicates), (2) the samples have equal volume, and (3) the samples have analytical recovery that is (on average) equal. Even if these estimates are unbiased, any random difference between the counts of indigenous (oo)cysts (x_1, x_2) will inflate the variability among matrix spike recovery estimates. Bias due to unequal sample volumes can easily be addressed by using the matrix spike recovery estimate $[N_{sp} - (V_{sp} / V_s) \cdot N_s] / T$, in which V_{sp} and V_s are the volumes of the spiked and unspiked samples. If there are N_s indigenous oocysts in an unspiked sample of volume V_s, then the number expected in an unspiked sample of volume V_{sp} is $V_{sp}/V_s \times N_s$. Discussion of Bayesian approaches to analyze individual matrix spike recovery estimates is included in Appendix B.2, but is not used in this thesis.

4.2.7 Internal Seed Recovery Estimates

Internal seed recovery estimates, such as those obtained using ColorSeed® (BTF Precise Microbiology, Inc.; Pittsburgh, PA), do not need to be corrected for the presence of indigenous particles (as in Section 4.2.6) because it is possible to distinguish between seeded particles from indigenous particles. In the model presented herein, it is assumed that the number of seeded particles is precisely known. If analytical recovery is regarded as a probability and all seeded particles are equally likely to be observed, then a binomial model can be used to describe analytical error as shown in Equation 4.13. This binomial model for internal seed recovery was also used by Petterson *et al.* (2007). If it is also assumed that recovery varies randomly among internally seeded samples according to a beta distribution, then the resulting joint distribution is shown in Equation 4.14. In these equations, n^* is the precisely known number of internal seed particles, x^* is the number of observed internal seed particles, p is the probability of recovery, and *a,b* are the shape parameters of the beta distribution for non-constant analytical recovery. The model with beta-distributed non-constant analytical recovery is equivalent to the beta-binomial recovery model presented in Section 4.2.2 because the presence of indigenous particles in the sample is overcome by using uniquely identifiable seeded particles. Use of internal seed recovery data to facilitate inference about unknown indigenous particle concentrations is addressed in Sections 5.1.3 and 6.1.1.

$$f\left(x^* \mid n^*, p\right) = \left[\frac{n^*!}{x^*!\left(n^* - x^*\right)!} p^{x^*} (1-p)^{n^*-x^*}\right] \tag{4.13}$$

$$f\left(x^*, p \mid n^*, a, b\right) = \left[\frac{n^*!}{x^*!\left(n^* - x^*\right)!} p^{x^*} (1-p)^{n^*-x^*}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}\right] \tag{4.14}$$

4.2.8 Critical Review of Other Probabilistic Recovery Models

Although beta distributions have often been used to model variability in analytical recovery among samples, consideration of seeding and analytical errors in recovery data analysis or use of probabilistic models such as those presented above is rare. Often, seeding error is ignored (*i.e.* the seeding methodology and uncertainty in the nominal seed doses are not described or considered) and recovery estimates are presented as though they are exact measurements of analytical recovery. In general, analysis of recovery data is limited to computation of a mean and standard deviation of the computed fractions of seeded particles that were observed, or basic hypothesis tests based upon these fractions. For example, the validation criteria for the initial precision and recovery test in Method 1623 (USEPA, 2005b) are based on the sample mean and sample relative standard deviation of four recovery estimates, and do not consider seeding error or the reduced analytical error associated with higher seed doses (as will be discussed in Section 4.4.2). In this section, several probabilistic recovery models that have been used for (oo)cyst recovery data are summarized and contrasted with the models presented above. Discussion of a probabilistic model addressing partial sample analysis and other errors in the enumeration of unseeded samples (Young and Komisar, 1999) and discussion of beta-Poisson dose-response models is also included because these models have important similarities to the recovery models presented herein.

Teunis *et al.* (1999) used beta-binomial models (such as the one presented in Section 4.2.2) for a variety of applications including viability of *Cryptosporidium* oocysts based on observed viable type morphology, recovery of *Cryptosporidium* oocysts, and removal of sulphite reducing clostridia or viruses by a water treatment process. Accordingly, analytical recovery was regarded as a probability. The example recovery dataset in Teunis *et al.* (1999) consisted of six samples (three each of treated and untreated water) seeded with varying quantities of *Cryptosporidium* oocysts (ranging from 2,478-

26,668). The numbers of seeded oocysts were not precisely known (they were seeded from a stock of well-known concentration), but this error was disregarded because the seed doses were high. The data were analyzed using a constant recovery model (*i.e.* only the binomial distribution with an unknown constant analytical recovery) and a variable recovery model identical to the beta-binomial model described herein. It was concluded that the recovered fractions varied considerably and that the beta-binomial model provided a superior fit. This result is not surprising, however, because two different water types were used (although this was not determined to have a significant effect), the seed doses were not precisely known, and the seed doses varied over a very large range (which could possibly result in non-random variations in analytical recovery). Furthermore, the beta-binomial model has two fitted parameters, while the constant recovery model has only one, and models with more fitted parameters generally yield better fits. Such a small and poorly controlled dataset cannot be used to evaluate the validity of the beta-binomial model. A large and well-controlled dataset is used herein in Section 4.4.2.

Petterson *et al.* (2007) used binomial and beta-binomial models for (oo)cyst recovery data and regarded analytical recovery as a probability. Specifically, a binomial model was used to model analytical error in internal seed recovery data (*e.g.* Section 4.2.7). This model was used when an internal seed recovery estimate was available for each of the environmental (oo)cyst samples, and there was no model for non-constant analytical recovery (*i.e.* each was considered independently without an overall distribution describing variability in analytical recovery among samples). If it is possible to describe the variability in analytical recovery among such samples using a distribution (*i.e.* if recovery varies randomly), then a superior analysis of the analytical recovery in each sample could be accomplished by using a non-constant analytical recovery distribution as well as the internal seed recovery data (*e.g.* Equation 4.14). Petterson *et al.* (2007) did not address this possibility, perhaps because the time series of non-replicate enumeration data that they considered may have non-random variations in analytical recovery (in which case the internal seed recovery estimate may be the only representative recovery information available). Use of informative beta-distributed priors for internal seed recovery in Bayesian analyses of enumeration data is addressed in Sections 5.1.3 and 6.3.1. When internal seed recovery data were available for only a few of the environmental samples, a beta-binomial model (*e.g.* Section 4.2.2) was used to describe variability among the available internal seed recovery estimates. It was assumed that there was no non-random variability in recovery among non-replicate samples and that the beta distribution was representative of the variability in recovery among the environmental samples for which no recovery data were collected. Further discussion of how these recovery models were incorporated into statistical analysis of indigenous (oo)cyst enumeration data is included in Sections 5.1.4 and 6.1.2.

Crainiceanu *et al.* (2003) developed a complex Bayesian model for recovery data collected in accordance with the United States' Information Collection Rule. The model accounted for several types of random errors that can affect recovery estimates as described herein (*e.g.* seeding error, partial sample analysis, and non-constant analytical recovery). Analytical recovery was typically regarded as a rate, although this rate was also defined as the probability of counting a microorganism. Seeding error was modelled using a gamma distribution that was fitted to a reported mean and variance of seed doses. It was noted that a continuous distribution could only be used for the discrete number of seeded particles because the seed doses were high (*i.e.* generally in the thousands). This seeding error model, however, is not suitable for low seed doses and the reasons for choosing the gamma distribution are not described. It appears that it was selected simply to enable fitting to the reported mean and variance of seed doses and that the associated fit was not evaluated. In contrast, choice of the Poisson distribution in Section 4.2.3 is based on relevant stochastic assumptions. On the other hand, the gamma distribution is quite versatile and is not restricted to the very specific assumptions associated with the Poisson model (*i.e.* it can possibly be used to model seeding errors that are more or less variable than the Poisson model). Fitted seeding error models are not addressed herein, but a binomial model may be appropriate for seed doses that are under-dispersed relative to a Poisson distribution (*e.g.* prepared seed doses with some variability), while a negative binomial model could be fitted to over-dispersed seed doses. Crainiceanu *et al.* (2003) modelled analytical error and partial sample analysis using a Poisson distribution in which the mean is equal to the product of the number of seeded particles, the rate of recovery, and the fraction of the original sample volume that was enumerated. As described in Sections 4.2.3 and 4.2.5, these errors should be described by a binomial model. The Poisson approximation to the binomial distribution is only valid if the number of seeded particles is large and either the recovery or the fraction of sample analyzed is small (Schmidt *et al.*, 2010b). If either assumption is untrue, then the Poisson model is inappropriate and can be substantially over-dispersed relative to the appropriate binomial model. Three alternative models (including a beta distribution) were used to account for random variability in the recovery rate (with consideration of possible covariates and random effects as described in Section 4.1.5). All three models restrained the recovery rates to values less than 1, which is more consistent with regarding recovery as a probability.

Teunis and Havelaar (1999) presented two probabilistic models for recovery estimates: a beta-binomial model for samples containing only seeded microorganisms (as in Teunis *et al.*, 1999) and a model for paired matrix spike samples. The matrix spike model that they developed is more complicated than the model presented in Section 4.2.6 because it addresses sets of matrix spike recovery data rather than each individual pair. It assumes that the concentration of indigenous microorganisms varies among the matrix spike pairs according to a gamma distribution, but is constant within pairs of samples. Likewise, it assumes that recovery varies among matrix spike pairs according to a beta distribution, but is constant within pairs of samples. These enhancements to the model presented herein may yield more informative results, but it is not clear how appropriate these assumptions are. For example, variability in indigenous microorganism concentration over time may not be gamma-distributed, or there may be temporal correlation between the pairs (*e.g.* the concentration during one matrix spike event may not be fully random if it is related to the concentration during the previous event). Additionally, the model for non-constant analytical recovery assumes that recovery is constant within pairs but varies among pairs. The model therefore describes non-random variations in recovery among non-replicates and assumes that there is no random variation in recovery among replicates. In general, the model has additional assumptions that are potentially less robust than the assumptions made herein, but it is likely more informative because it can be used to analyze entire matrix spike datasets rather than individual pairs. The model's applicability may also be increased by incorporating seeding error (as considered herein).

Young and Komisar (1999) developed probabilistic models for enumeration of environmental samples with partial sample analysis. Although the models do not address partial sample analysis in recovery experiments (as described in Section 4.2.5), the models are sufficiently similar to warrant discussion in this section. In the first model, they assumed that the number of (oo)cysts in a subsample withdrawn from a well-mixed resuspended pellet is Poisson-distributed. As described in the Sections 4.2.3 and 4.2.5, a binomial model is much more appropriate for this sub-sampling error. A Poisson approximation to the binomial model is only valid if the number of (oo)cysts in the sample is high and the fraction of resuspended pellet that is sub-sampled is low. The second model assumes that the sub-sampling error is negative binomially distributed because clumping has been alleged as a cause of data that are over-dispersed with respect to the Poisson distribution in a variety of types of water. This model may also have been motivated by high variability in counts of repeated subsamples from a well-suspended pellet. This negative binomial model is inappropriate for three main reasons: (1) it is generally infeasible to obtain sample-specific information (*i.e.* parameters for the negative binomial model) about clumping in the resuspended pellet, (2) statistically valid inference about (oo)cyst concentrations from subsamples that are inadequately homogenized is not possible, and (3) the binomial model for subsamples from well-mixed suspensions suggests that the sub-sampling error is typically under-dispersed relative to the Poisson distribution while the negative binomial model is necessarily over-dispersed. Like the negative binomial model, the binomial model has two parameters; however, these parameters are more easily accommodated because one is the unknown parameter of interest (the actual number of particles in the sample) and the other is the measurable fraction of sample that is enumerated. The parameters in the negative binomial model correspond to the expected number of (oo)cysts in the subsample (which is essentially the product of the two parameters in the binomial model) and the dispersion of (oo)cysts in the resuspended pellet (which is virtually immeasurable). Inference about the actual concentration of particles represented by a sample is complicated when enumeration includes sub-sampling from an inadequately homogenized resuspended pellet because the error is non-random. For example, if four (oo)cysts are enumerated in a subsample, it is not possible to infer how many (oo)cysts were actually present (*e.g.* a clump of four all captured in the subsample, or a higher number of which only four were obtained) without sample-specific data demonstrating the representativeness of the subsamples. Clumping or other non-random (oo)cyst distributions in sub-sampled resuspended pellets is an important consideration worthy of further research and modelling, but does not justify the use of stochastically inappropriate or practically infeasible models.

Although the applications differ, the beta-Poisson recovery model presented in Section 4.2.3 is stochastically equivalent to the beta-Poisson dose-response model that is often used in Quantitative Microbial Risk Assessment (*e.g.* Furumoto and Mickey, 1967; Haas, 1983; Regli *et al.*, 1991; Haas *et al.*, 1993; Haas *et al.*, 1996; Teunis and Havelaar, 1999; Teunis and Havelaar, 2002; Smeets *et al.*, 2007). In the beta-Poisson dose-response model, it is assumed that pathogens are randomly distributed in drinking water with a known homogeneous concentration so that the number of pathogens in a volume of water ingested by a consumer is Poisson-distributed. This random sampling error is comparable to Poisson-distributed seeding error. It is then assumed that there is a host-microorganism interaction represented by a probability that each individual pathogen will survive to initiate infection. Of the pathogens that are ingested, the number that survive to initiate infection is binomially distributed. Accordingly, the binomial and Poisson distributions can be combined into a single Poisson distribution with the mean equal to the expected number of pathogens in the sample multiplied by the probability of each pathogen surviving to initiate infection. Finally, it is assumed that the host-microorganism interaction varies among consumers according to a beta distribution. The resulting beta-Poisson dose-response model is stochastically equivalent to Equation 4.6 (except that it is used to describe the probability of infection rather than the distribution of the number of pathogens that initiate infection). Furumoto and Mickey (1967) developed an approximation of the beta-Poisson dose-response model that has subsequently been reparametrized as Equation 4.15 (Haas, 1983), in which $P*$ is the probability of a consumer becoming infected, N is the average dose (the product of pathogen concentration and ingested drinking water volume), and α,β are the parameters of the beta distribution. This approximation is used because the actual probability, as shown in Equation 4.16, cannot be evaluated explicitly. The approximation assumes that α is much smaller than β and that β is much larger than one (Teunis and Havelaar, 2002). The fitted parameter values obtained using this approximation for various dose-response datasets are often incompatible with these assumptions, which results in a dose-response model that is not stochastically beta-Poisson (Teunis and Havelaar, 1999). Figure 4.3 shows the actual and approximated beta-Poisson models obtained using the rotavirus beta-Poisson dose-response model parameters reported in Regli *et al.* (1991). Although the approximation error is sometimes small and may be insignificant relative to other errors and uncertainties in these risk assessment models, it is inappropriate to continue using faulty simplifying approximations given the considerable advances in computation over the past half-century.

$$P* = 1 - \left[1 + \frac{N}{\beta}\right]^{-\alpha} \tag{4.15}$$

$$P* = 1 - \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 e^{-Np} p^{\alpha-1} (1-p)^{\beta-1} \, dp \tag{4.16}$$

4.3 STRATEGIES TO IMPROVE DESIGN OF RECOVERY EXPERIMENTS

When a sample is seeded with a known quantity of analyte particles, the fraction of the seeded particles that are observed is an estimate of analytical recovery (with some modification necessary for matrix spike recovery estimates as discussed in Sections 4.1.7 and 4.2.6). The actual value of analytical recovery associated with each sample, whether regarded as a probability or as a rate, is unknown. The number of seeded particles that are observed in a sample is affected by seeding error and analytical error as well as non-constant analytical recovery (as modelled in Section 4.2). As the

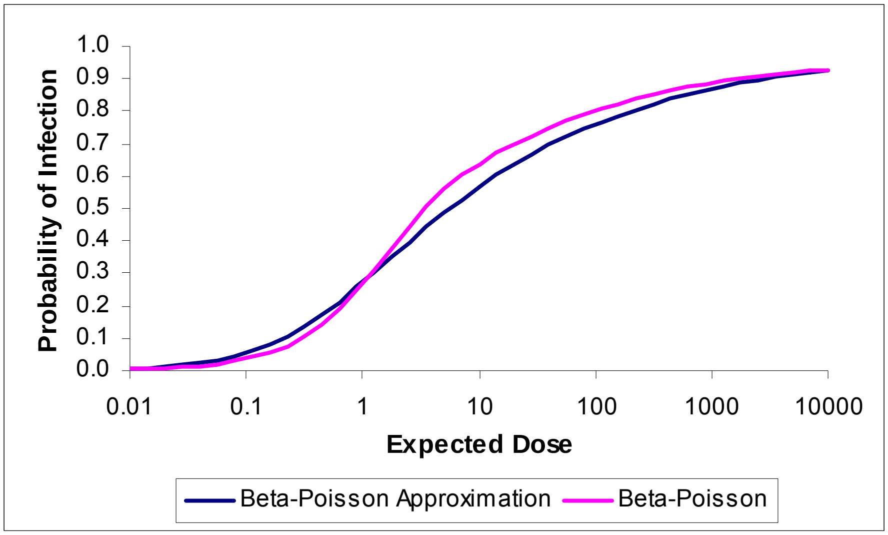


Figure 4.3: Approximated and actual beta-Poisson dose-response models

These dose-response models, based on reported rotavirus model parameters $\alpha = 0.26$, $\beta = 0.42$ (Regli *et al.*, 1991) show that the commonly used beta-Poisson approximation can be inconsistent with the actual beta-Poisson model when the assumptions are not met.

effects of these errors upon the variability of counts in fully replicated recovery data increases, the precision of the associated recovery estimates decreases. The distribution of recovery estimates will, on average, be more variable than analytical recovery itself because of these random errors.

This section addresses the design of recovery experiments with fully replicated seeded samples (*i.e.* equivalent seed dose and seeding process, equivalent methodology, and no non-random variations in analytical recovery among samples). Such experiments are commonly conducted to investigate factors that affect analytical recovery or to compare variations in methodology when a method that is known to have low or highly variable recovery is being used. These experiments may also be used for method validation: examples include the initial precision and recovery experiment in Method 1623 (USEPA, 2005b), and seeding experiments used to demonstrate the recovery (or sensitivity) of new methods (*e.g.* Teixeira *et al.*, 2007). Finally, recovery experiments that are representative in both sample characteristics and methodology of the enumeration of indigenous particles can be used to quantify the recovery associated with unseeded samples (Nahrstedt and Gimbel, 1996; Schmidt *et al.*, 2010a). The objective in such experiments is to quantify the mean and variance (or the distribution) of analytical recovery without the confounding effects of seeding error and analytical error (*i.e.* to quantify non-constant analytical recovery alone). The factors that affect the precision of recovery estimates in these fully replicated recovery experiments will similarly affect the precision of recovery estimates in other applications (*e.g.* matrix spike and internal seed recovery estimates).

Monte Carlo simulations (Section 3.3.1) are used herein to investigate the effects of experimental design (*e.g.* the quantity of seeded particles, the precision of the seed doses, and the number of fully replicated samples) upon estimates of the mean and standard deviation of analytical recovery (*i.e.* the sample mean and sample standard deviation of the recovery estimates). Many decisions are based upon these values including the results of many hypothesis tests that are used to compare alternative methodologies or to evaluate the impacts of factors that may affect analytical recovery. The validation criteria of Method 1623 (USEPA, 2005b), for example, specify a minimum average recovery and a maximum relative standard deviation. Monte Carlo simulation of recovery data using any of the models presented in Section 4.2 (given random number generation algorithms for the uniform, binomial, Poisson, beta, and gamma distributions as well as hypothetical seed doses and non-constant analytical recovery distribution parameters) is a relatively simple task. Recovery parameters (*a*,*b* or α,β) from preliminary results or similar methods can be used to evaluate a particular recovery experiment, or a wide range of these parameters can be used to obtain general guidelines for recovery experiment design (Schmidt *et al.*, 2010a). The simulations presented in this section use the beta-binomial recovery model (Section 4.2.2, Equation 4.3) when the seed dose is assumed to be precisely known, the beta-Poisson recovery model (Section 4.2.3, Equation 4.6) when samples are seeded with an aliquot from a stock of known concentration, and the beta-binomial partial sample analysis model (Section 4.2.5, Equation 4.10). The analyses presented herein were completed in Microsoft Excel^{TM} using the Visual Basic Editor. Random number generation algorithms for the various distributions can be found in Ahrens and Dieter (1974) or modified for Basic from *Numerical Recipes in C* (Press *et al.*, 1992).

To evaluate the accuracy and precision of estimates of the mean and standard deviation of analytical recovery, the results of a single hypothetical recovery experiment (*e.g.* a specified number of samples seeded in a particular way) can be simulated multiple times to generate a distribution of estimated means and standard deviations. These can then be compared to the actual mean and standard deviation of analytical recovery associated with the selected beta or gamma non-constant analytical recovery distribution. Herein, each hypothetical recovery experiment design was simulated 10,000 times and the results (*e.g.* the estimated means and standard deviations) were summarized by 95% probability intervals (with 2.5% of the simulated values excluded from each tail of the distribution). These probability intervals from multiple hypothetical experimental designs can be used to investigate changes in experimental design that may yield more accurate and/or precise estimates of the mean and standard deviation of analytical recovery (Schmidt *et al.*, 2010a).

The analyses presented herein use four sets of beta distribution parameters that represent the permutations of two means (0.7518, 0.2529) and two standard deviations (0.0435, 0.0854). The associated beta distributions are plotted in Figure 4.4. For each set of parameters, probability intervals based on the Monte Carlo simulations were calculated for all permutations of the following experimental design parameters: (1) seed doses of 50, 100, and 500 particles, (2) numbers of fully replicated samples from 3 to 20, and (3) seed doses assumed to be precisely known or Poisson-distributed. The resulting probability intervals for both the mean and standard deviation of the recovery estimates for all of these various experimental designs are plotted in Figures 4.5, 4.6, 4.7, and 4.8. Each figure corresponds to a set of beta distribution parameters and includes a table with four panels (the probability intervals for the mean and standard deviation are on the left and right respectively, and the probability intervals for known and Poisson-distributed seed doses are on the top and bottom respectively).

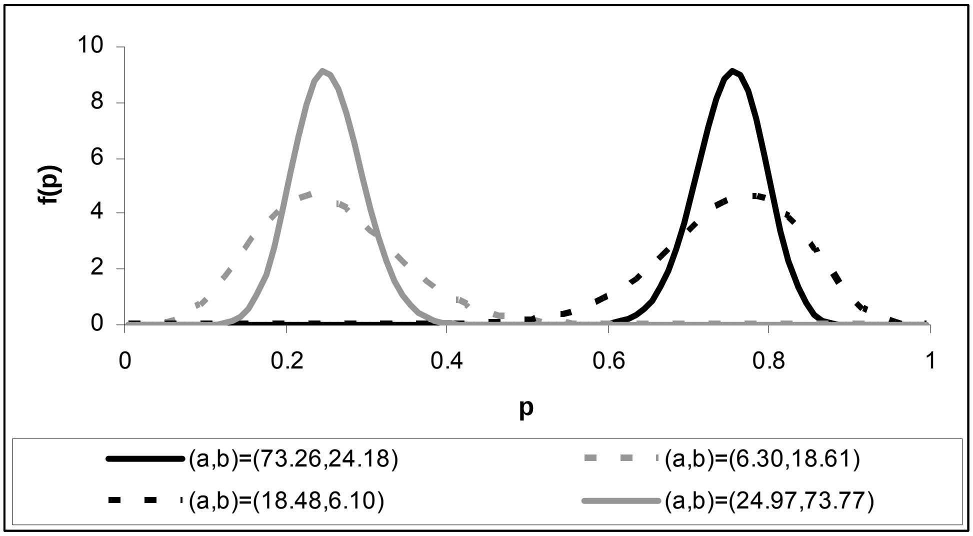


Figure 4.4: Probability density functions of example recovery distributions
(Schmidt *et al.*, 2010b)

In general, these figures show that more precise estimates of the mean and standard deviation of analytical recovery can be obtained by increasing the number of fully replicates samples, by increasing the number of seeded particles, and by seeding precisely known numbers of particles. They also demonstrate that the number of observed particles divided by the presumed number of seeded particles is an unbiased estimate of analytical recovery because the probability intervals for the mean of the recovery estimates are centred upon the actual mean analytical recovery. Mathematical proof that these recovery estimates are unbiased is included in Section 8.1.1. Each of these results is predictable based on relatively simple theory. For example, the standard error of the mean of any population with finite variance is $\sigma/\sqrt{n}$, which shows that more precise estimates of the mean can be obtained by increasing the number of samples (n).

Increasing the number of seeded particles (if analytical recovery is independent of seed dose) will result in the fraction of seeded particles that are observed being a more precise estimate of the actual probability of recovery. This has two effects: it results in more precise estimates of the mean recovery (as indicated by narrower probability intervals) and it results in more accurate and more precise estimates of the standard deviation of recovery. The probability intervals illustrate that the estimated standard deviation becomes more precise as the number of seeded particles in increased because the intervals become narrower. Similarly, it is illustrated that the estimated standard deviation becomes more accurate because the average of the standard deviation estimates converges upon the actual standard deviation of recovery. The latter result is most apparent in the known seeding standard deviation plots in Figures 4.5 and 4.8 because the 50-particle probability intervals do not include the actual value of the standard deviation of analytical recovery. The sample standard deviation of the recovery estimates is biased high because of analytical error, and analytical error is greatest when the number of particles is small. The effect of this error is relatively large in Figures 4.5 and 4.8 (which

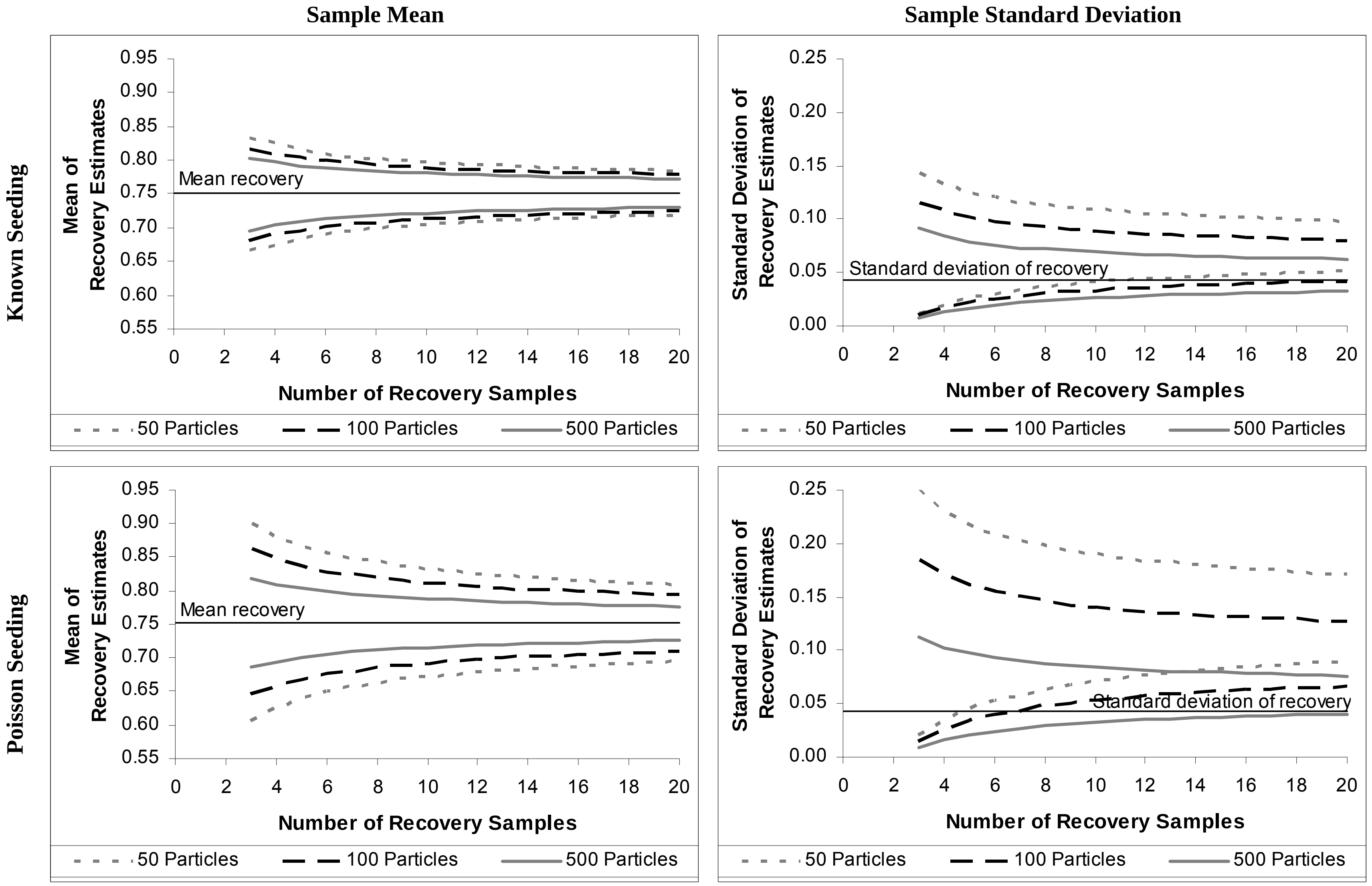


Figure 4.5: 95% probability intervals for 3-20 samples with known or Poisson seeding of 50-500 particles (*a*,*b*) = (73.26, 24.18)
(Schmidt *et al.*, 2010b)

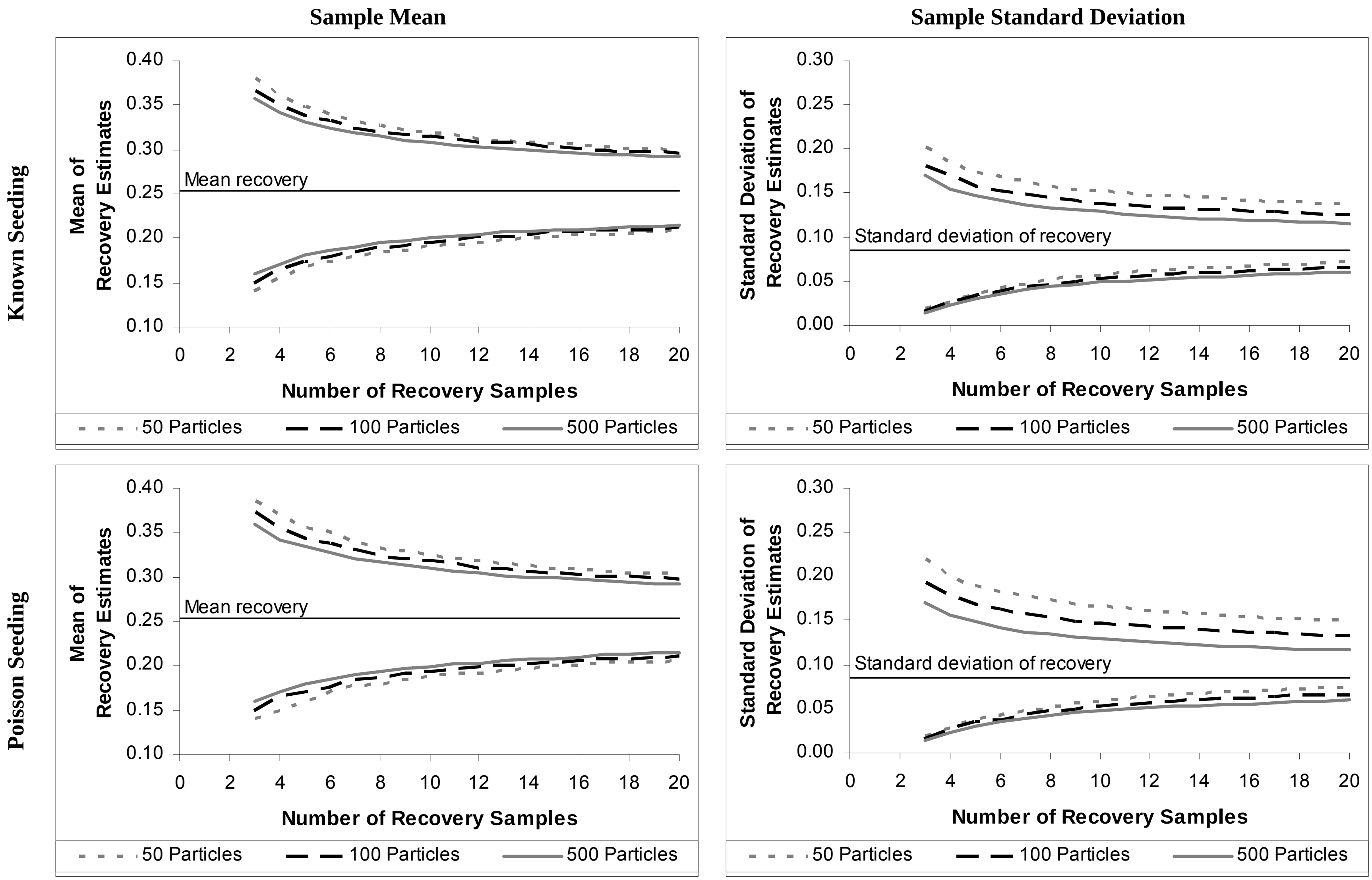


Figure 4.6: 95% probability intervals for 3-20 samples with known or Poisson seeding of 50-500 particles (*a*,*b*) = (6.30, 18.61)

(Schmidt *et al.*, 2010b)

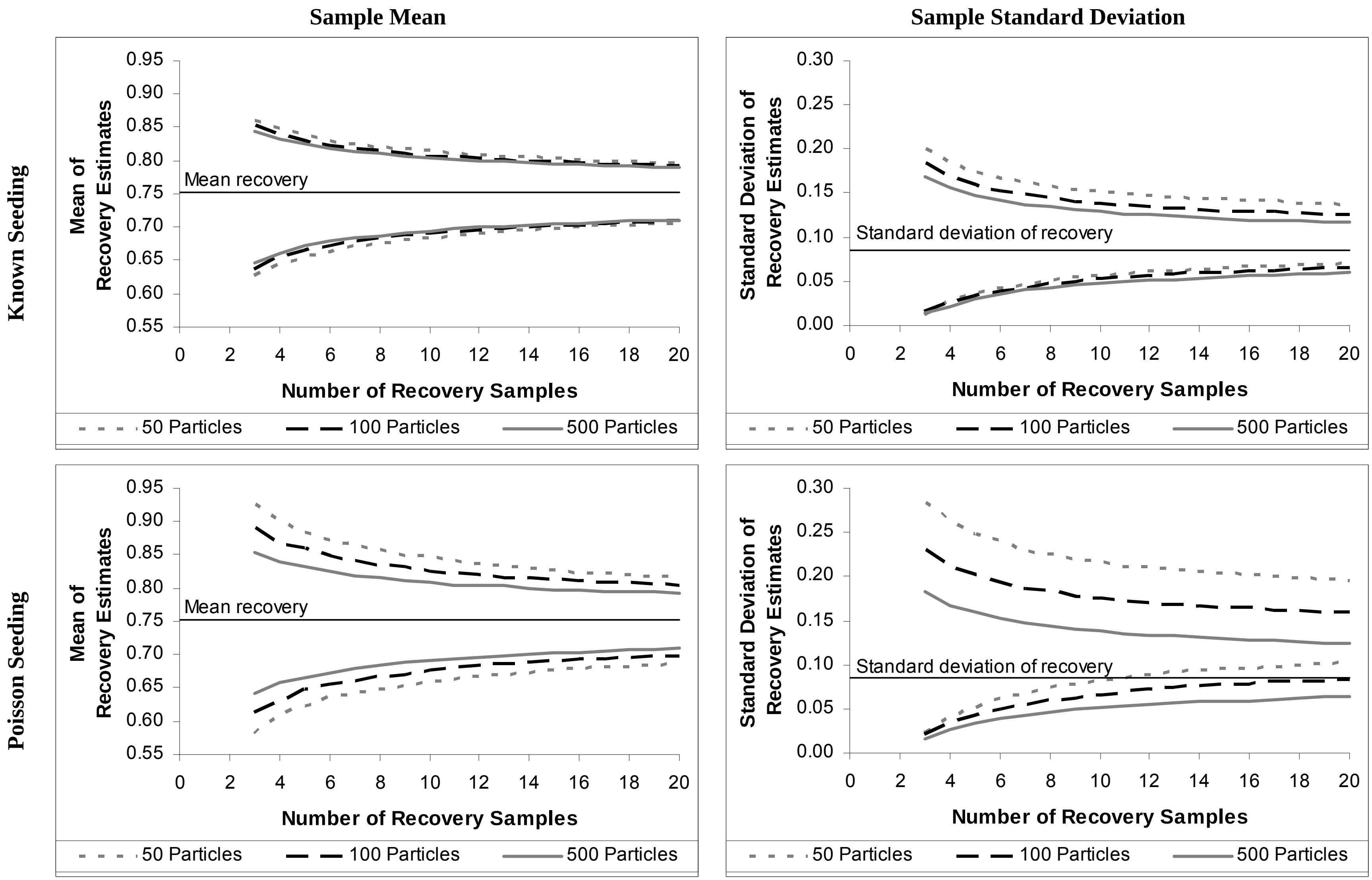


Figure 4.7: 95% probability intervals for 3-20 samples with known or Poisson seeding of 50-500 particles (*a*,*b*) = (18.48, 6.10)
(Schmidt *et al.*, 2010b)

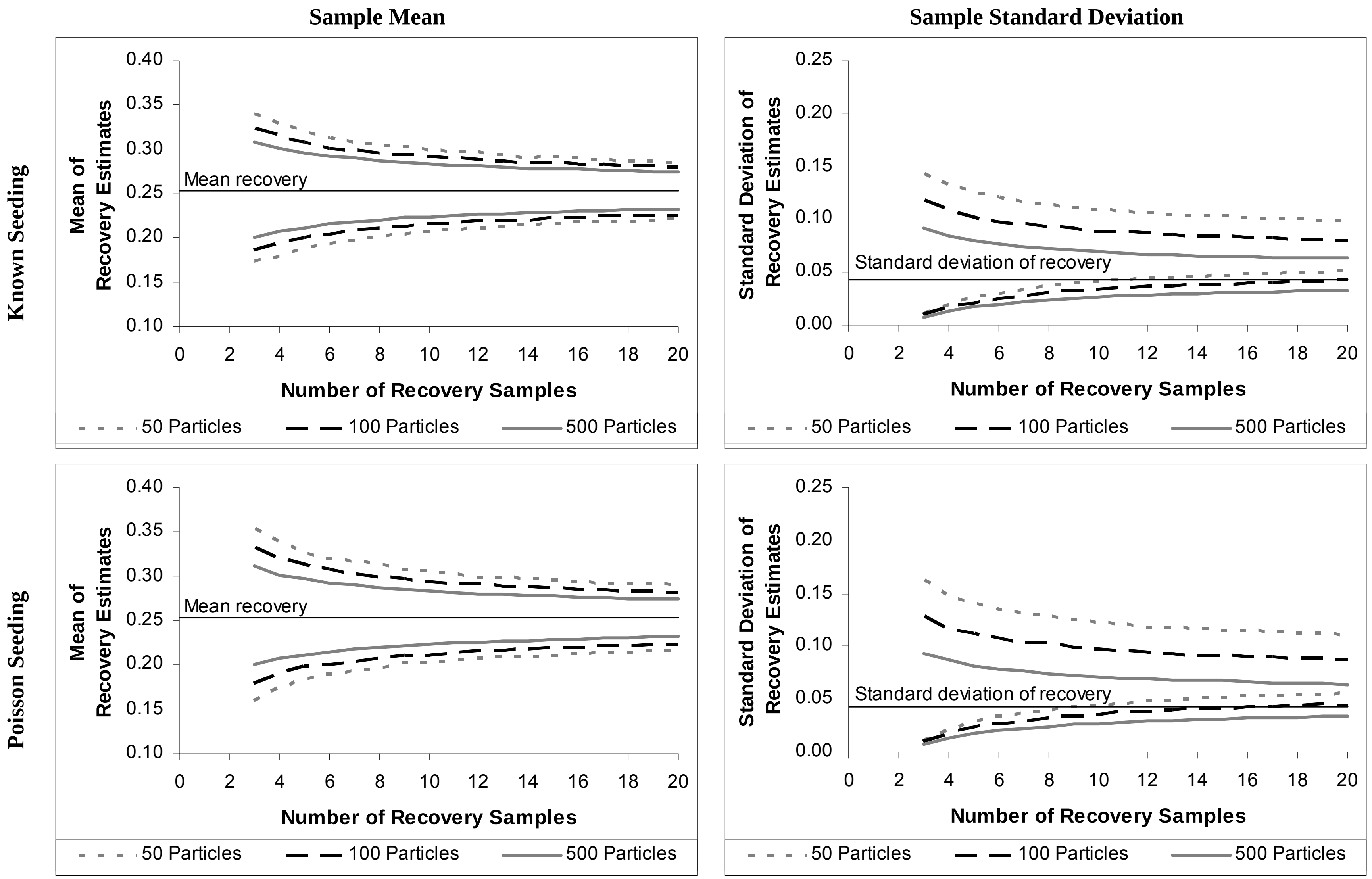


Figure 4.8: 95% probability intervals for 3-20 samples with known or Poisson seeding of 50-500 particles (*a,b*) = (24.97, 73.77)
(Schmidt *et al.*, 2010b)

have a low actual standard deviation of analytical recovery) and is relatively small in Figures 4.6 and 4.7 (which have a higher actual standard deviation of analytical recovery). This demonstrates that using larger numbers of seeded particles in most impactful when the method has relatively consistent recovery because the additional variability due to analytical error is muted when recovery is highly variable. For any seeding approach and non-constant analytical recovery distribution, there becomes a quantity of seeded particles beyond which further increases in the quantity of seeded particles has a negligible impact upon the precision of the recovery estimates (*i.e.* the probability intervals suggest that the estimated standard deviation is unbiased and can only be further narrowed by increasing the number of fully replicated samples). This happens when the excess variability due to seeding and analytical error becomes small relative to the variability in analytical recovery itself (as discussed in Section 8.1.4). Accordingly, this quantity of seeded particles beyond which recovery estimates can be regarded as precise is dependent upon the variability in analytical recovery and the seeding error.

Using imprecise seed doses will result in variability in the recovery data that is attributable to seeding error rather than non-constant analytical recovery. This result is demonstrated in all of the figures because the probability intervals associated with Poisson-distributed seeding are wider than those associated with precisely known seed doses (which indicates that the recovery estimates are less precise) and because the standard deviation of the recovery estimates is an even more biased estimate of the standard deviation of analytical recovery. The effects of using Poisson-distributed seed doses rather than precisely known seed doses are most apparent in Figures 4.5 and 4.7 (which have higher mean analytical recovery). Explanation of this phenomenon is addressed in Section 8.1.4.

The effect of partial sample analysis upon the precision of recovery estimates must also be considered. For example, if analytical recovery is independent of seed dose (n or λ) and of partial sample analysis (θ), it may be of interest to know if enumerating 10% of a sample containing 100 particles yields as accurate of a recovery estimate as enumerating an entire sample seeded with 10 particles. These two types of samples have the same 'effective seed dose' of 10 particles, but the sub-sampling errors associated with partial sample analysis must be considered. The following analysis assumes (as discussed in Section 4.2.5) that the samples are well-mixed at the time of sub-sampling. In the models with Poisson-distributed seeding, partial sample analysis has a trivial effect upon the precision of recovery estimates. If a sample is seeded with λ particles and only a fraction (θ) of the sample is enumerated (*e.g.* Equation 4.11), then the resulting count is as variable as a completely enumerated sample with a seed dose of $\lambda\theta$ (*e.g.* Equation 4.6). The effective seed dose in each of these scenarios is Poisson-distributed with an average of $\lambda\theta$, so there is no difference in the precision of the associated recovery estimates. If, on the other hand, a sample is seeded with a precisely known number of particles, then partial sample analysis will result in excess variability because the number of particles in the sample (n) is precisely known while the number of particles in the enumerated portion of sample is binomially distributed with mean $n\theta$. This 'effective seeding error' due to partial sample analysis will at least partially negate the benefits of using precisely known seed doses (Schmidt *et al.*, 2010b). Further discussion of the effects of partial sample analysis upon the precision of recovery estimates is addressed in Section 8.1.3.

Figure 4.9 provides an example of the effects of partial sample analysis upon recovery estimates. The recovery estimates used in the beta-binomial recovery model are $x/n\theta$. The first row of panels illustrates an example in which the effective seeding error associated with enumerating 10% of each 100-particle sample is nearly equivalent to Poisson-distributed seeding with an average of 10 particles per sample. This is because the binomial distribution describing the number of particles in the subsample is well-approximated by a Poisson distribution when the number of particles is large and the enumerated subsample is a small fraction of the sample. This is also an example in which samples are seeded with precisely known numbers of seeded particles and recovery estimates exceeding 100% (due to partial sample analysis) are commonly obtained. The second row of panels illustrates an example in which the effective seed dose in each sample is 50 particles. The results of enumerating 50% of each 100-particle sample are between the results obtained by precisely known or Poisson-distributed seeding of 50 particles per sample. In general, partial sample analysis of precisely seeded samples resulting in an effective seed dose of $n\theta$ will result in recovery estimates that are more precise than entirely enumerated samples with the same Poisson-distributed effective seed dose and less precise than entirely enumerated samples with the same precisely known effective seed dose. As described above, the difference between these two extremes can be trivial when seeding error contributes minimally to overall variability in recovery estimates. The results will tend towards precisely known seeding with the same effective seed dose for large values of θ.

4.4 QUANTITATIVE ANALYSIS OF REPLICATE RECOVERY DATA

The variability in analytical recovery among replicates may be of interest in method development and comparison studies, method validation, or in experiments that are used to predict the analytical recovery associated with unseeded samples. Often the mean of the recovery estimates is presented along with a standard deviation or a standard error. The mean and the standard deviation of the recovery estimates may also be used in a hypothesis test to determine whether the difference in mean analytical recovery between two methods can be classified as statistically significant (*e.g.* whether the observed difference is random or due to an actual difference in mean analytical recovery). Such tests are fraught with problems because (1) the recovery estimates are more variable than analytical recovery itself due to seeding and analytical errors (as demonstrated in Section 4.3), (2) analytical recovery is unlikely to be normally distributed (because it is necessarily positive, it may also be necessarily less than 100%, and it often follows a skewed distribution), and (3) recovery estimates with different quantities of seeded particles that are regarded as replicates will have unequal errors. These many limitations of commonly used statistical approaches demonstrate a need for more robust strategies to interpret recovery data. Furthermore, strategies to appropriately interpret data from the enumeration of unseeded samples (*e.g.* Chapter 5) require information about the analytical recovery of the enumeration method. Accordingly, it is necessary to develop models that describe random variation in analytical recovery among samples (*i.e.* non-constant analytical recovery) such as the beta or gamma distributions used in Section 4.2.

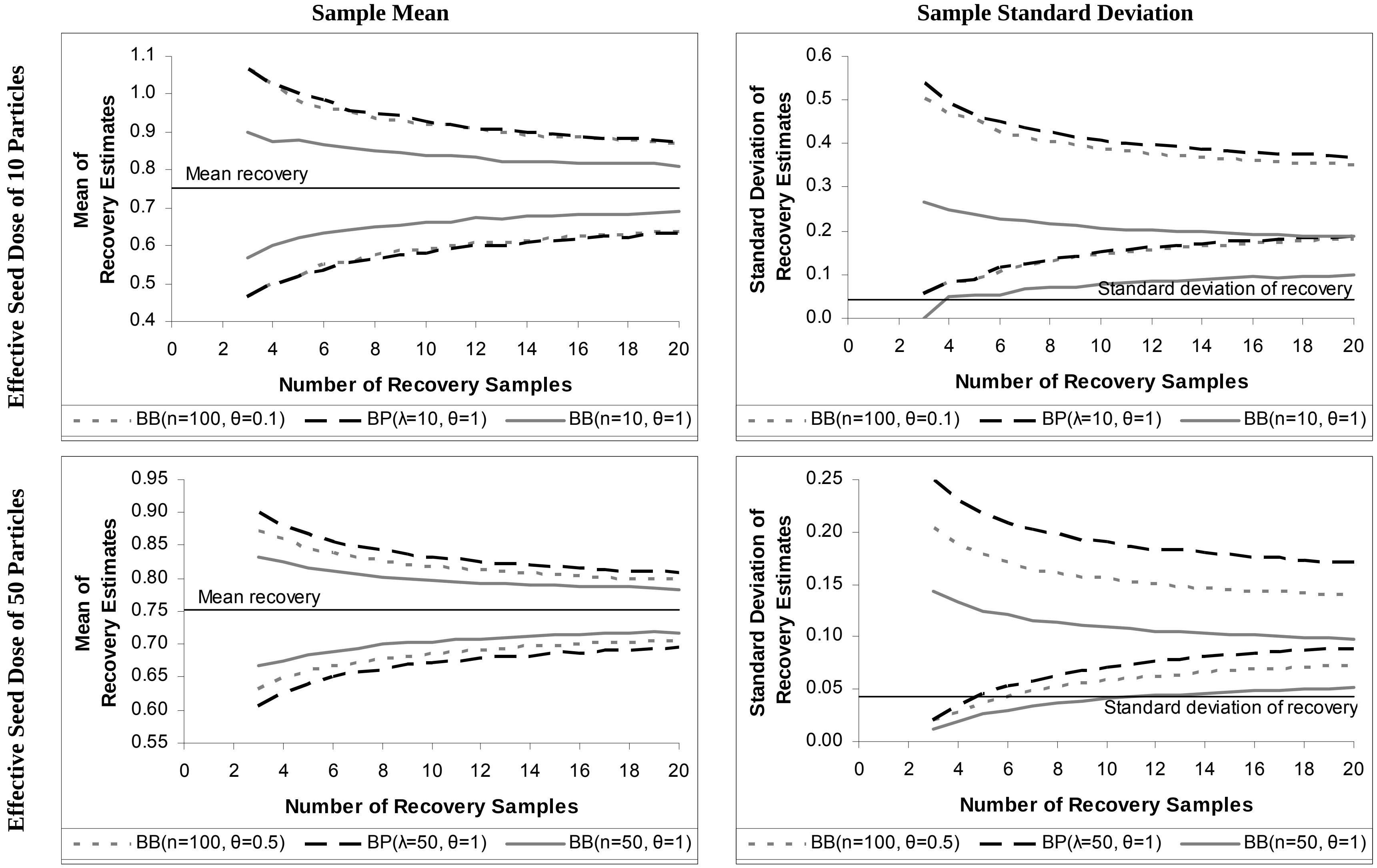


Figure 4.9: Effective seeding error due to partial sample analysis

95% probability intervals for 3-20 samples, (*a*,*b*) = (24.97, 73.77), BB = beta-binomial recovery model, BP = beta-Poisson recovery model.

This section commences with discussion of common strategies that are used to obtain point estimates of the parameters of the beta or gamma distributions that describe non-constant analytical recovery (Section 4.4.1). Discussion of the Method 1623 (USEPA, 2005b) initial precision and recovery experiment validation criteria and analysis of a large ongoing precision and recovery dataset follows in Section 4.4.2. In Section 4.4.3, Bayes' theorem is used as a tool to enable quantitative description of uncertainty in the parameters of the non-constant analytical recovery distributions. Gibbs sampling is used to analyze a case study dataset in Section 4.4.4. This section is concluded (Section 4.4.5) with a summary of the limitations of classical statistical tools (such as confidence intervals and hypothesis tests) that may be used to analyze recovery data and with a discussion of the potential value of Bayesian analyses as a more robust alternative.

4.4.1 Fitting a Distribution to Recovery Estimates

Many probabilistic models have been developed to incorporate information about analytical recovery into analyses of microorganism or discrete particle enumeration data. In general, these models have assumed that recovery varies randomly among samples (and perhaps also non-randomly in models with potential covariates such as Crainiceanu *et al.* (2003)). Exceptions include Parkhurst and Stern (1998), who did not address the variability in analytical recovery among samples, and a model presented in Petterson *et al.* (2007) that excluded variability in analytical recovery among samples because each count of indigenous (oo)cysts was paired with an internal seed recovery estimate. The models that address variability in analytical recovery among samples using beta distributions have typically obtained parameter estimates using the method of moments (*e.g.* Nahrstedt and Gimbel, 1996), maximum likelihood estimation (*e.g.* Teunis *et al.*, 1999; Teunis and Havelaar, 1999; Pouillot *et al.*, 2004; Schmidt *et al.*, 2010a), or Bayesian methods (*e.g.* Crainiceanu *et al.*, 2003; Petterson *et al.*, 2007; Schmidt *et al.*, 2010a). This section discusses parameter estimation using the method of moments and maximum likelihood estimation. Bayesian methods are discussed in Section 4.4.3.

The method of moments (Section 3.1.3) has been used by Nahrstedt and Gimbel (1996) to fit a beta distribution describing variability in analytical recovery among samples to the mean and variance of recovery estimates. Their approach used a biased variance estimate that divided the sum of the squared deviations from the sample mean by the number of data (while the unbiased estimate divides by one less than the number of data). The method of moments estimators used to fit a beta or gamma distribution directly to the recovery estimates (Schmidt *et al.*, 2010a) are presented in Equations 4.17 and 4.18 respectively. In these equations, $\hat{p}$ is the recovery estimate (*i.e.* x/n or x/λ), $\bar{x}_{\hat{p}}$ is the sample mean of the recovery estimates, and $s_{\hat{p}}$ is the sample standard deviation of the recovery estimates.

$$\hat{a} = \bar{x}_{\hat{p}} \left(\frac{\bar{x}_{\hat{p}} \left(1 - \bar{x}_{\hat{p}}\right)}{s_{\hat{p}}^2} - 1 \right), \quad \hat{b} = \frac{\left(1 - \bar{x}_{\hat{p}}\right)}{\bar{x}_{\hat{p}}} \hat{a} \tag{4.17}$$

$$\hat{\alpha} = \frac{\bar{x}_{\hat{p}}^2}{s_{\hat{p}}^2}, \quad \hat{\beta} = \frac{s_{\hat{p}}^2}{\bar{x}_{\hat{p}}} \tag{4.18}$$

Parameter estimation using this particular method of moments approach can, however, be quite inappropriate because it assumes that each recovery estimate is an exact measurement of analytical recovery. It has been demonstrated in Section 4.3 that the standard deviation of recovery estimates can be a substantially biased estimate of the true standard deviation of analytical recovery due to the effects of seeding and analytical error, which are not addressed in this method of moments approach. Discussion of method of moments estimators that account for seeding and analytical error is included in Section 8.1.5, but these have been found to be inappropriate because they can yield infeasible parameter estimates. A final problem with the method of moments is that samples that are seeded with unequal quantities of particles or that are seeded with inconsistent precision will yield recovery estimates with unequal measurement error (*i.e.* some recovery estimates may be more informative than others). This issue is difficult to address in method of moments parameter estimation approaches.

Maximum likelihood estimation (Section 3.1.3) is a more robust approach to obtain parameter estimates. Depending upon the probabilistic model that is used in the likelihood function, it can easily account for seeding and analytical error and for non-replicate recovery estimates. Maximum likelihood can be used for non-replicate data because it is not strictly necessary for each sample to have the same seed dose. In fact, beta-binomial and beta-Poisson recovery data (that are representative of the same non-constant analytical recovery distribution) can be combined. It is assumed throughout this section that all recovery estimates are independent; the analytical recovery of one sample is unlikely to affect the analytical recovery of another if all samples are processed using identical methodology and if there is no cross-contamination.

Like the method of moments, using maximum likelihood estimation to fit beta or gamma distributions directly to recovery estimates is typically incorrect because seeding and analytical errors in the recovery estimates are not addressed (*i.e.* each estimate is assumed to be an exact measurement of analytical recovery). This approach is explored herein to illustrate the effects of ignoring seeding and analytical errors relative to other likelihood approaches that are based on more appropriate probabilistic models. The likelihood functions for the 'beta-only' and 'gamma-only' recovery models are presented in Equations 4.19 and 4.20 respectively (Schmidt *et al.*, 2010a). The associated maximum likelihood estimates are the values of the parameters that maximize the likelihood function on the feasible space (*e.g.* $a > 0$, $b > 0$ or $\alpha > 0$, $\beta > 0$). These likelihood functions can be differentiated using the digamma function $\Psi(.)$ to represent the derivative of the gamma function $\Gamma(.)$; however, the critical points cannot be found explicitly. Maximum likelihood estimates were computed in Microsoft Excel$^{\text{TM}}$ using the sum of the sample-specific log-likelihoods (*i.e.* the natural logarithm of the likelihood of each sample, summed across all samples) and the Solver optimization tool. Proof that these likelihood functions are unimodal is shown in Appendix A.

$$L(a,b) = \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)^{r} \left(\prod_{i=1}^{r} \hat{p}_i\right)^{a-1} \left(\prod_{i=1}^{r} (1-\hat{p}_i)\right)^{b-1} \tag{4.19}$$

$$L(\alpha,\beta) = \left(\frac{1}{\beta^{\alpha}\Gamma(\alpha)}\right)^{r} \left(\prod_{i=1}^{r} \hat{p}_i\right)^{\alpha-1} e^{-\sum_{i=1}^{r}\hat{p}_i \big/ \beta} \tag{4.20}$$

A more appropriate maximum likelihood estimation approach can be developed by using a probabilistic model that is more representative of the process by which the available data were generated. The likelihood functions associated with the beta-binomial (Section 4.2.2), beta-Poisson (Section 4.2.3), and negative binomial (Section 4.2.4) recovery models are presented in Equations 4.21, 4.22, and 4.23 respectively. In each of these equations, the term in square brackets is a constant and can be dropped from the likelihood calculations if optimization is the sole objective. In the following analyses, however, comparison of likelihoods obtained using the various models requires these terms to be included. The likelihood function for the beta-Poisson model has been simplified from Equation 4.6 by integrating the probability of recovery (*p*) out of the model: however, the integral cannot be solved explicitly and the likelihood function must be numerically approximated. Further discussion of this numerical approximation is included in Section 5.3.1. Modification of these likelihood functions for partial sample analysis (*i.e.* using the probabilistic models presented in Section 4.2.5) is not discussed herein, but has been presented in Schmidt *et al.* (2010b). If a recovery dataset includes samples with precisely known and Poisson-distributed seed doses that are assumed to be representative of the same non-constant analytical recovery distribution, then the likelihood for the samples with precisely known seed doses (Equation 4.21) can be multiplied by the likelihood for the samples with Poisson-distributed seed doses (Equation 4.22) to yield an overall likelihood function. Maximum likelihood estimates were computed using the sum of the log-likelihoods and the Solver optimization tool in Microsoft Excel$^{\text{TM}}$. It has not been proven herein that these likelihood functions are unimodal, but no evidence has been found to the contrary.

$$L(a,b)=\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)^{r}\prod_{i=1}^{r}\frac{\Gamma(x_i+a)\Gamma(n_i-x_i+b)}{\Gamma(n_i+a+b)}\cdot\left[\prod_{i=1}^{r}\frac{n_i!}{x_i!(n_i-x_i)!}\right] \tag{4.21}$$

$$L(a,b)=\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)^{r}\prod_{i=1}^{r}\int_0^1 e^{-\lambda_i p}p^{x_i+a-1}(1-p)^{b-1}\,dp\cdot\left[\prod_{i=1}^{r}\frac{\lambda_i^{x_i}}{x_i!}\right] \tag{4.22}$$

$$L(\alpha,\beta)=\frac{1}{(\Gamma(\alpha))^{r}}\prod_{i=1}^{r}\frac{\Gamma(x_i+\alpha)\beta^{x_i}}{(\lambda_i\beta+1)^{x_i+\alpha}}\cdot\left[\prod_{i=1}^{r}\frac{\lambda_i^{x_i}}{x_i!}\right] \tag{4.23}$$

If analytical recovery does not vary among samples, then the variability in recovery estimates is exclusively due to seeding and analytical error. The likelihood functions for the precisely known and Poisson-distributed seed doses would be represented by Equation 4.24 and 4.25 respectively (the ‘binomial-only’ and ‘Poisson-only’ recovery models). These likelihood functions are easily differentiable and yield the maximum likelihood estimates $\Sigma x_i/\Sigma n_i$ and $\Sigma x_i/\Sigma\lambda_i$ respectively.

$$L(p)=(p)^{\sum x_i}(1-p)^{\sum(n_i-x_i)}\cdot\left[\prod_{i=1}^{r}\frac{n_i!}{x_i!(n_i-x_i)!}\right] \tag{4.24}$$

$$L(p)=e^{-p\sum\lambda_i}p^{\sum x_i}\cdot\left[\prod_{i=1}^{r}\frac{\lambda_i^{x_i}}{x_i!}\right] \tag{4.25}$$

To demonstrate and examine the methods presented herein, a recovery dataset (Table 4.4) is analyzed using each of the parameter estimation approaches presented above (Schmidt *et al.*, 2010a). Such an analysis is for illustrative purposes only, because only the model that best represents the measurement errors in the data should be used. This would not be possible if some of the seed doses were not integers or some of the recovery estimates were greater than 100% because the beta-binomial recovery model would be incompatible with the data. Vesey *et al.* (1993) did not provide sufficient details on seeding methodology to determine whether the seed doses can be assumed to be precisely known or Poisson-distributed, so it is unclear which model is most appropriate. It is assumed in the following analyses that the three types of water did not have an impact upon analytical recovery and that the data can be regarded as nine replicates. The point estimates obtained using each of the nine parameter estimation approaches described above are summarized in Table 4.5 along with the associated log-likelihoods (where applicable).

Table 4.4: Example *Cryptosporidium parvum* oocyst recovery data (Vesey *et al.*, 1993)

Sample	Deionized Water			Tap Water			River Water		
Seeded Number	608	608	608	608	608	608	608	608	608
Observed Number	472	485	431	420	468	458	420	479	481

Table 4.5: Non-constant analytical recovery distribution point estimates
(modified from Schmidt *et al.*, 2010a)

Parameter Estimation Approach & Probabilistic Model	Distribution Parameters	Mean Recovery	Std. Dev. of Recovery	log-Likelihood
Method of Moments – Beta	(73.26, 24.18)[a]	0.751827	0.043535	-
Max. Likelihood – 'Beta-only'	(85.62, 28.27)[a]	0.751778	0.040302	16.168563
Max. Likelihood – Beta-binomial	(104.53, 34.51)[a]	0.751795	0.036503	-41.544792
Max. Likelihood – Beta-Poisson	(287.08, 94.76)[a]	0.751840	0.022076	-41.755720
Method of Moments – Gamma	(298.24, 0.002521)[b]	0.751827	0.043535	-
Max. Likelihood – 'Gamma-only'	(328.93, 0.002286)[b]	0.751827	0.041454	15.887204
Max. Likelihood – Neg. Binomial	(1236.16, 0.000608)[b]	0.751827	0.021384	-41.775075
Max. Likelihood – 'Binomial-only'	0.751827[c]	0.751827	-	-53.878639
Max. Likelihood – 'Poisson-only'	0.751827[c]	0.751827	-	-42.018656

[a] beta distribution parameters (a,b); [b] gamma distribution parameters (α,β); [c] constant analytical recovery (p)

The parameter estimates associated with each model indicate a mean recovery near 75.18%. Similar means are expected when the data represent fully replicated analyses because each datum is equally informative. Differences may occur, however, if some samples are seeded with different quantities of seeded particles because some recovery estimates will be more precise than others. The two method of moments approaches as well as the 'beta-only' and 'gamma-only' maximum likelihood approaches assume that all recovery estimates are equally informative. In contrast, the other approaches will implicitly give greater weight to the recovery estimates that are less impacted by seeding and analytical error. This is because the sample-specific likelihood function (*i.e.* the likelihood function obtained for a single datum) for more precise recovery estimates is narrower, which causes the overall likelihood to be governed more so by the precise recovery estimates than the imprecise recovery estimates (*e.g.* as shown in Figure 4.10).

The standard deviation values associated with the different sets of parameter estimates in Table 4.5 show considerable differences. This is because of the ways in which the various models handle seeding and analytical errors. The recovery estimates are assumed to be exact measurements of analytical recovery in the two method of moments approaches and the 'beta-only' and 'gamma-only' maximum likelihood approaches. Accordingly, these approaches resulted in the highest standard deviations because they attributed all of the variability among the recovery estimates to non-constant analytical recovery when some of the variability is likely due to seeding and analytical error. The beta-binomial maximum likelihood approach yields a smaller standard deviation because it accounts for the impacts of analytical error upon the recovery estimates. The beta-Poisson and negative binomial maximum likelihood approaches yield the smallest standard deviations because they account for seeding and analytical errors. These results demonstrate that it is imperative to use a model that appropriately addresses the seeding and analytical errors in the recovery estimates. If recovery estimates are collected in such a way that the effects of seeding and analytical error are demonstrated to be small (*e.g.* using the Monte Carlo approaches presented in Section 4.3), then the recovery estimates will be relatively precise and the various parameter estimation approaches that assume

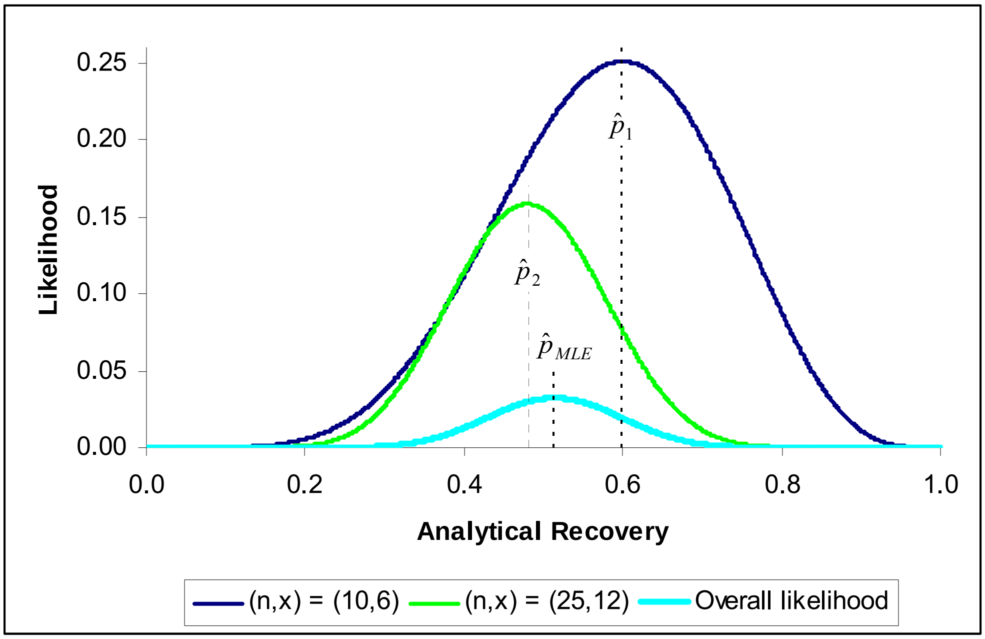


Figure 4.10: Effect of the precision of recovery estimates upon likelihood functions
The likelihood functions associated with two independent binomially distributed counts (assuming that analytical recovery is constant but unknown) are shown along with the overall likelihood function that is obtained when the two data are considered together. In general, the overall likelihood function is higher when the likelihood associated with the second (more precise) sample is higher. The overall maximum likelihood estimate ($\hat{p}_{MLE}$) is closer to the recovery estimate obtained from the more precise sample ($\hat{p}_2$) than the recovery estimate from the less precise sample ($\hat{p}_1$).

non-constant analytical recovery may yield similar model distributions. Nonetheless, maximum likelihood estimates obtained using models that appropriately represent the seeding and analytical errors in the recovery data are most appropriate (even if the impacts of these errors are small).

The log-likelihood values associated with the various maximum likelihood estimates can be used to investigate the fit of the various models. In general, a higher log-likelihood indicates a better fit if the models are directly comparable. The ‘beta-only’ and ‘gamma-only’ models are directly comparable with each other but not with the other models (because the random variables are continuous recovery values rather than discrete counts). The likelihoods indicated that the ‘beta-only’ model provides a slightly better fit to the recovery estimates than the ‘gamma-only’ model, but the difference is minimal as shown in Figure 4.11. The other five models are directly comparable with each other. The fitted beta-binomial, beta-Poisson, and negative binomial models have similar log-likelihoods (*i.e.* they provide similar fit), but the interpretation of the models with respect to random errors in the data is different. The fitted ‘binomial-only’ model has a much lower log-likelihood than the beta-binomial recovery model, which indicates that there is excess variability in the data (due to non-constant analytical recovery or unaddressed seeding error). The fitted ‘Poisson-only’ model has only slightly lower log-likelihood than the beta-Poisson or negative binomial recovery models, which indicates that adding a distribution for non-constant analytical recovery to the model (and an extra fitted parameter) yields only a slightly better fit. Models with more fitted parameters often yield a better fit, so only substantial differences in likelihood between the ‘binomial-only’ or ‘Poisson-only’ models and the beta-binomial, beta-Poisson, or negative binomial models are important.

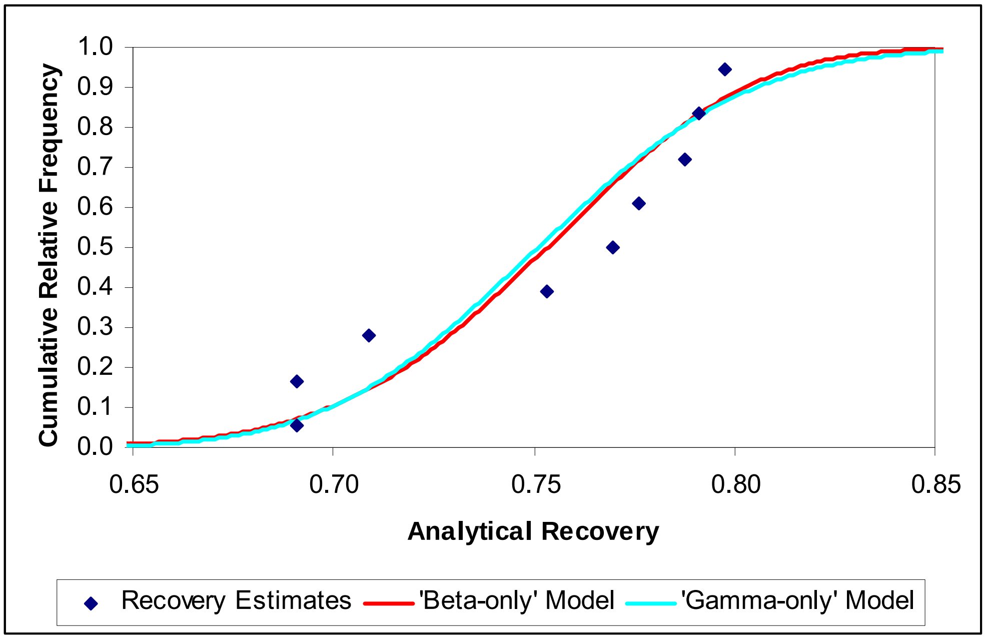


Figure 4.11: Fit of the ‘beta-only’ and ‘gamma-only’ models to the recovery estimates

It is common for models to be selected based only upon good fit without appropriate consideration to the associated mathematical assumptions and random errors being modelled. One problem with such an approach is that it is somewhat arbitrary: equivalent experiments could be fitted with different models and the variability in the data would arbitrarily be assigned to different sources. For example, the superior fit off the beta-binomial recovery model relative to the 'binomial-only' model could arise from non-constant analytical recovery or from seeding error that is falsely attributed to non-constant analytical recovery. This maximum likelihood analysis of the Table 4.4 data using the various models cannot be used to select which model is most appropriate for the data. Such a decision can only be based upon careful consideration of the seeding errors in the recovery data and the assumptions of the various models. Accordingly, recovery experiments and recovery estimates that are published without details about seeding (*i.e.* the presumed seed doses and the possible errors therein) have minimal scientific value: it is not possible to carry out a correct statistical analysis of a recovery dataset without appropriate consideration of measurement errors in the recovery estimates. Therefore, it is recommended that the seeding method should always be explicitly described, the presumed seed doses should be stated, and raw count data should be provided where possible.

4.4.2 Case Study: Analytical Recovery of Method 1623 for Reagent Water Samples

Method 1623 (USEPA, 2005b) is known to yield imperfect analytical recoveries of *Cryptosporidium* oocysts and *Giardia* cysts. The extensive quality control program associated with this method includes an initial precision and recovery (IPR) experiment and ongoing precision and recovery (OPR) analyses using reagent water samples seeded with specified quantities of oocysts and cysts. This section discusses statistical considerations associated with the IPR validation criteria and provides a case study analysis of OPR recovery data.

Statistical considerations in the IPR validation criteria

The IPR experiment consists of four reagent water samples seeded with 100-500 oocysts and 100-500 cysts. The mean and relative standard deviation (RSD) of the recovery estimates must meet the following validation criteria: the mean of the recovery estimates must be between 24-100% and the RSD ($s/\bar{x}\cdot 100\,\%$) for *Cryptosporidium* and *Giardia* must be below 55% and 49% respectively. These validation criteria have several limitations as listed below.

- The standard error of the mean of four recovery estimates ($\sigma/2$) is relatively large. A high or low mean recovery can be obtained by chance (*e.g.* as demonstrated by the probability intervals for the mean of the recovery estimates in Section 4.3) because the estimated mean recovery is imprecise. This validation criterion may not effectively discriminate against laboratories with low analytical recovery and may inadvertently discriminate against some laboratories with acceptable analytical recovery.
- The RSD is a quotient of two imprecise values. A randomly high RSD can result when a randomly high sample standard deviation is paired with a randomly low sample mean. Likewise, a randomly low RSD can result from a randomly low sample standard

deviation and a randomly high sample mean. This validation criterion does not effectively discriminate against laboratories with unacceptably variable analytical recovery and may inadvertently discriminate against laboratories with well-controlled analytical recovery.

- The sample standard deviation of the recovery estimates that is used to calculate the RSD is biased high due to seeding and analytical errors (which are ignored in the validation criteria). Accordingly, more precise estimates of the mean recovery and values of the RSD that are lower (on average) can be obtained by seeding samples with high and precisely known numbers of (oo)cysts.

Figures 4.12 and 4.13 illustrate these problems for a hypothetical laboratory that has an actual mean analytical recovery of 30% and an actual RSD of 40% (*e.g.* beta-distributed analytical recovery with $a = 4.08$, $b = 9.51$). Figure 4.12 shows the distribution of sample means from 10,000 simulated IPR experiments for each of four seeding scenarios (100 or 500 particles, precisely known or Poisson-distributed seeding). The 24% validation criterion is shown in each panel along with the percent of simulated IPR experiments in which the validation criterion was not met. Between 11.8-13.8% of the simulated experiments did not meet the mean recovery validation criterion, and experiments in which the number of seeded particles was lower and/or not precisely known were slightly less likely to meet this criterion (because the sample mean is less precise as discussed in Section 4.3). Figure 4.13 shows the distribution of the sample relative standard deviations for the same simulations along with the 55% validation criterion for *Cryptosporidium* relative standard deviation. Between 13.6-20.1% of the simulated IPR experiments did not meet the RSD validation criterion. The RSD validation criterion was less likely to be met in experiments in which the seed dose was low and/or not precisely known (because the sample mean and sample standard deviation are less precise and because the sample standard deviation is more likely to be biased high, as shown in Section 4.3). The fraction of simulated IPR experiments in which at least one of the *Cryptosporidium* validation criteria was not met ranged from 26.1-32.8% and followed a similar pattern. As illustrated in Section 4.3, increasing the seed dose or switching from Poisson-distributed to precisely known seed doses has the greatest effect upon the estimated mean and (relative) standard deviation of the recovery estimates when the variability in analytical recovery is low. Therefore these observed effects of changing the seed dose or seeding methodology will be greater when analytical recovery is less variable and can become negligible when analytical recovery is highly variable.

In Method 1623, the validation criteria must be met for both *Cryptosporidium* and *Giardia*. There are effectively four validation criteria that must be met simultaneously, while Method 1622 (USEPA, 2005a) has two validation criteria for *Cryptosporidium* alone. As described herein, a laboratory that meets the validation criteria on average may randomly fail one or more of the validation criteria because the sample mean and sample (relative) standard deviation of four recovery estimates are imprecise. The validation criteria are collectively conservative because acceptable laboratories have a high chance of failing one or more of the validation criteria and unacceptable laboratories are unlikely to randomly meet all four criteria.

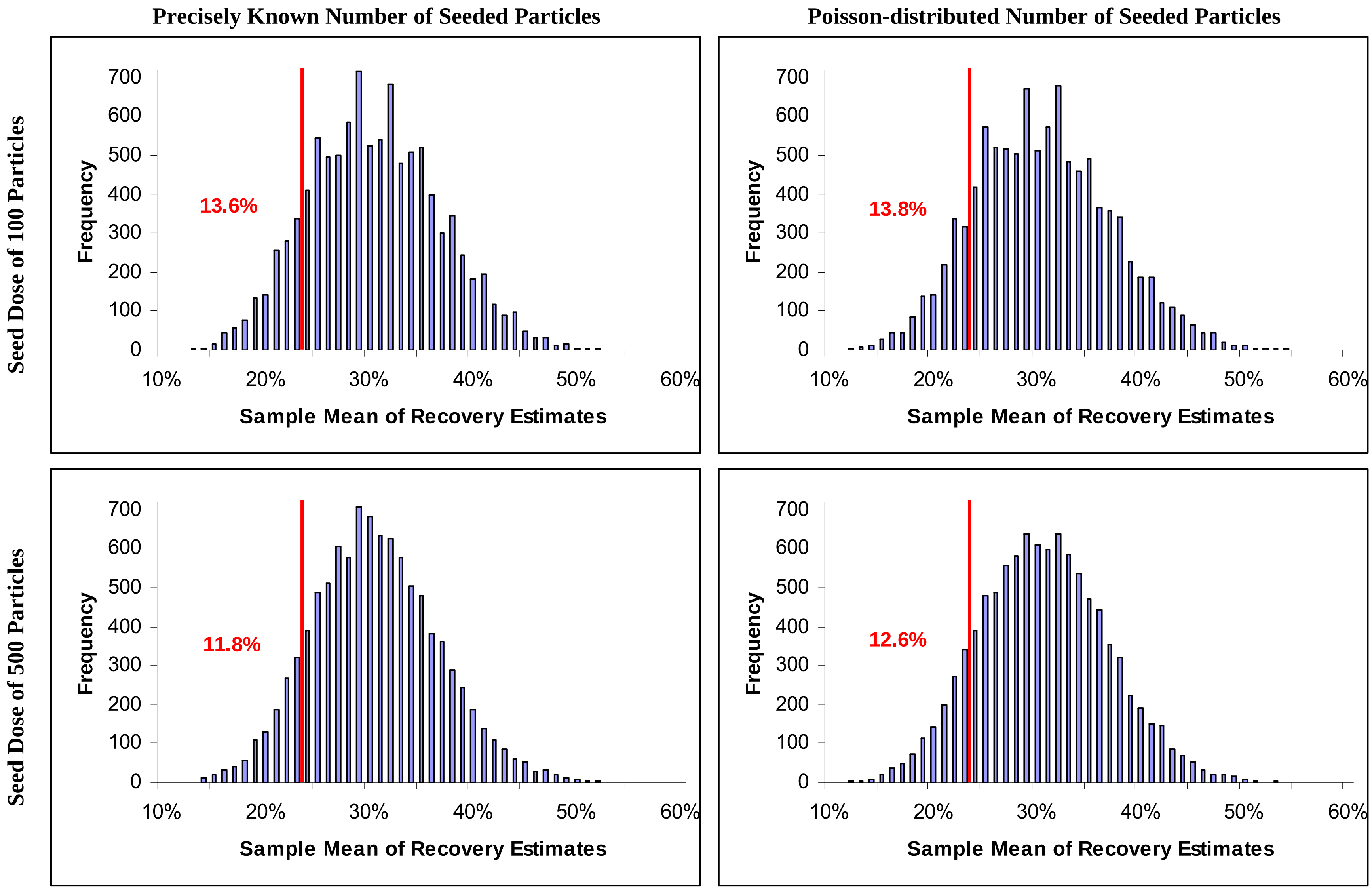


Figure 4.12: Initial precision and recovery experiment sample means

10,000 simulated experiments with beta distribution parameters (a,b) = (4.08, 9.51) corresponding to an actual mean analytical recovery of 30%

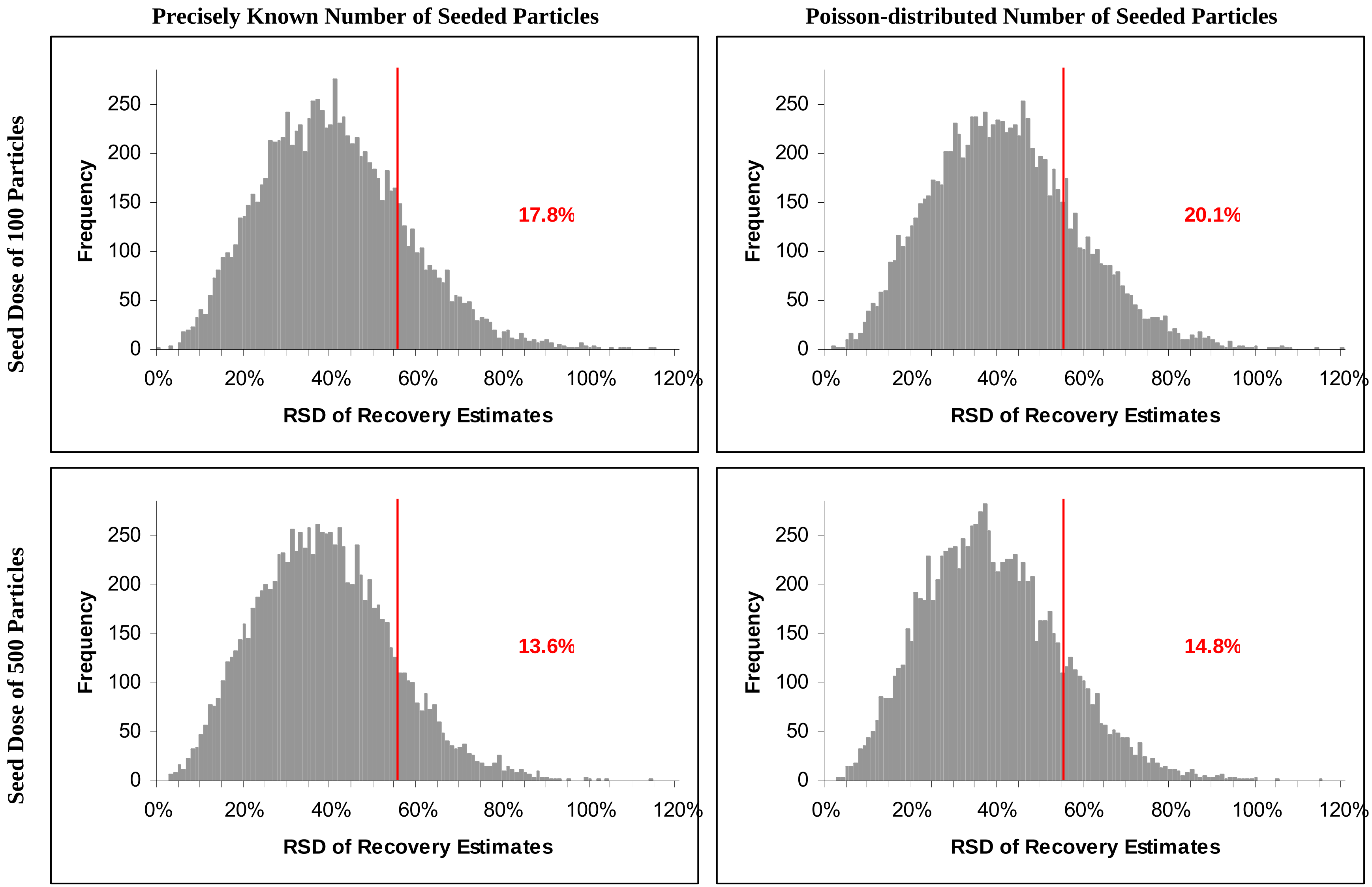


Figure 4.13: Initial precision and recovery experiment relative standard deviations

10,000 simulated experiments with beta distribution parameters (*a*,*b*) = (4.08, 9.51) corresponding to an actual relative standard deviation of 40%

It is concluded that the existing IPR validation criteria may not appropriately discriminate between laboratories with acceptable and unacceptable analytical recovery because laboratories with acceptable analytical recovery may have a high probability of failing to meet the validation criteria and laboratories can reduce the variability of their IPR recovery estimates by using larger and more precise seed doses. More appropriate validation criteria that account for seeding and analytical errors in the IPR recovery data and that consider the uncertainty in the estimated mean and relative standard deviation of analytical recovery (perhaps using the statistical analyses presented in Section 4.4.3) should be adopted to discriminate more appropriately between laboratories (and/or proposed variations in methodology) that have acceptable and unacceptable analytical recovery. At a minimum, the seeding requirements (*i.e.* seed dose and seeding method) should be more strictly controlled so that validation results cannot be manipulated by changing the quantity of seeded (oo)cysts and the precision therein.

Analysis of a large OPR recovery dataset

Ongoing precision and recovery analyses are completed to ensure continuous compliance with quality control criteria. OPR analyses are required at least once per week (when Method 1623 is being used) or at least once per 20 samples if more than 20 samples are analyzed in a week. The validation criterion for each OPR recovery estimate is 11-100% for *Cryptosporidium* and 14-100% for *Giardia*. A laboratory that processes many (oo)cyst samples over a long period of time will build up an extensive OPR dataset. An OPR dataset (Appendix D.1) from an anonymous Method 1623 approved laboratory is used herein to investigate the distribution of analytical recovery in reagent water and to look for possible temporal trends in the recovery data. The dataset consists of 444 OPR samples collected over several years, each of which was seeded with EasySeed® (BTF Precise Microbiology, Inc.; Pittsburgh, PA). As discussed in Section 4.1.1, there is slight variability among EasySeed doses that is indicated by a mean and standard deviation for the batch of doses. It is assumed herein that the number of seeded particles as reported by the laboratory is precisely known (these vary from 98-101 oocysts and 98-101 cysts per sample). Point estimates for the parameters of the beta distribution describing non-constant analytical recovery were obtained for oocysts and cysts using the beta-binomial maximum likelihood approach presented in Section 4.4.1. The parameter estimates were a = 11.51, b = 14.12 (which corresponds to a mean and a standard deviation of 44.9% and 9.64% respectively) for *Cryptosporidium* and a = 12.82, b = 15.91 (which corresponds to a mean and standard deviation of 44.6% and 9.12% respectively) for *Giardia*.

To evaluate the fit of the beta-binomial model to the data, the frequency of specific (oo)cyst counts is compared with the frequency predicted by the model. The presented model frequency accounts for the different seed doses in the OPR samples by calculating the sample-specific beta-binomial probability for each possible count (as a function of the presumed seed dose) and summing these probabilities across all of the OPR samples. Figure 4.14 shows the frequency (top panels) and cumulative frequency (bottom panels) for the (oo)cyst counts. The beta-binomial model describes the oocyst count data reasonably well (although the model seems slightly low in some areas and slightly high in others). Most of the discrepancy between the actual and modelled frequency in the histogram probably arises from the large number of bins and small number of data. The random deviation between relative frequencies and probabilities is only expected to become small when the number of data in each bin is much larger than the frequencies shown (which range from zero to 21). The beta-binomial model appears to yield a slightly better fit for the *Giardia* data than for the *Cryptosporidium* data. These results seem to support use of the beta-binomial recovery model because the model fits the data reasonably well. It is not clear from this analysis, however, if there are any trends in the mean and or variability of analytical recovery. If there are trends, then analytical recovery may not be independent and identically distributed throughout the dataset.

To investigate these OPR data further, Figure 4.15 shows the recovery estimates as a function of the chronological sample index. The two top panels show the 95% probability interval for the recovery estimates based on the beta-binomial model using the maximum likelihood parameter estimates and a seed dose of 99 (oo)cysts. These intervals represent a range within which 95% of all recovery estimates would be expected to fall if the model is correct. The two lower panels show a control chart with a moving sample mean computed from the most recent 20 samples (or fewer at the beginning). Control limits (representing the moving mean +/- two times the similarly calculated moving standard deviation) are also plotted. These plots are based on a template obtained from the U.S. EPA[2]. This type of graphic representation of continued laboratory performance is required in Method 1623 (Section 9.7.6 of the regulation). These control charts can help to display trends in the data (*e.g.* changes in the mean and or variability of the recovery estimates). For example, it is relatively clear that OPR oocyst recovery has increased gradually over time with higher variability at the beginning and end of the dataset than in the middle. While the beta-binomial model may provide a reasonable fit to the data, it will over-estimate variability in analytical recovery because some of the variability is due to the gradually increasing mean. The recovery at the end of the dataset, for example, is somewhat higher and less variable than what would be estimated using the entire dataset. The *Giardia* dataset, in contrast, shows a relatively consistent mean and standard deviation. It seems reasonable to assert that these recovery estimates are independent and identically distributed (because there are no apparent trends), which also suggests that analytical recovery itself is independent and identically distributed. If the seed doses were not well-controlled (*i.e.* varying doses and/or varying precision), then trends in the variability of the recovery estimates observed in such plots could arise from seeding and/or analytical error rather than non-constant analytical recovery.

Figure 4.16 shows a scatter plot of the 444 pairs of OPR oocyst and cyst recovery estimates, which are apparently uncorrelated. If *Cryptosporidium* oocysts and *Giardia* cysts shared many of the same analytical errors (*e.g.* losses in a particular stage of sample processing or counting errors during enumeration), then the results would likely show stronger correlation (although correlation could be obscured by seeding and analytical error). This poor correlation illustrates a need for investigation of the factors that affect recovery of (oo)cysts in the method and research into whether or not seeded (oo)cysts are good surrogates for recovery of indigenous (oo)cysts.

[2] Laboratory Quality Assurance Evaluation Program for Analysis of *Cryptosporidium* under the Safe Drinking Water Act (http://www.epa.gov/safewater/disinfection/lt2/pdfs/labs_approval/labs_examplecontrolchart.xls, accessed 16/06/2010)

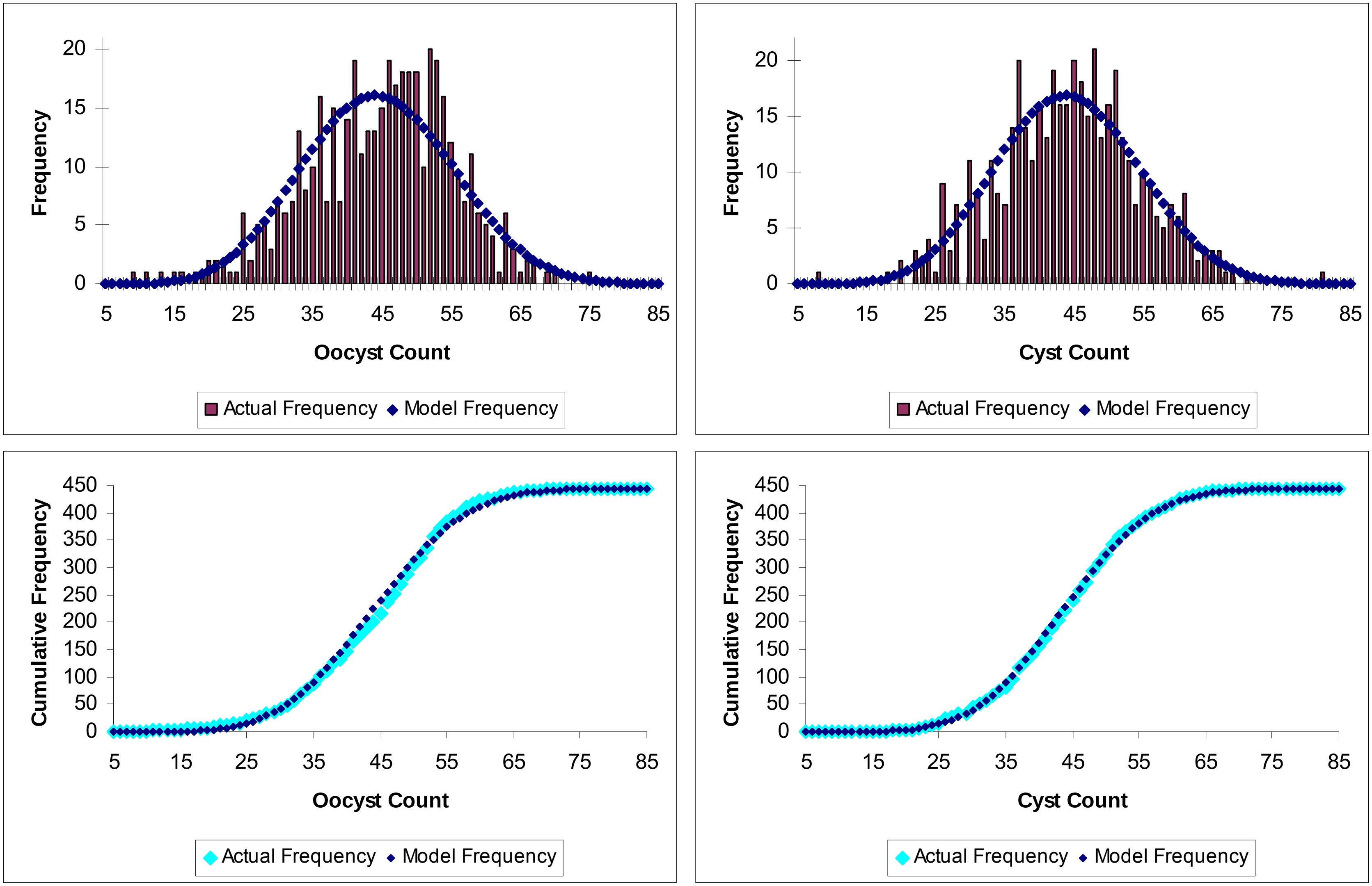
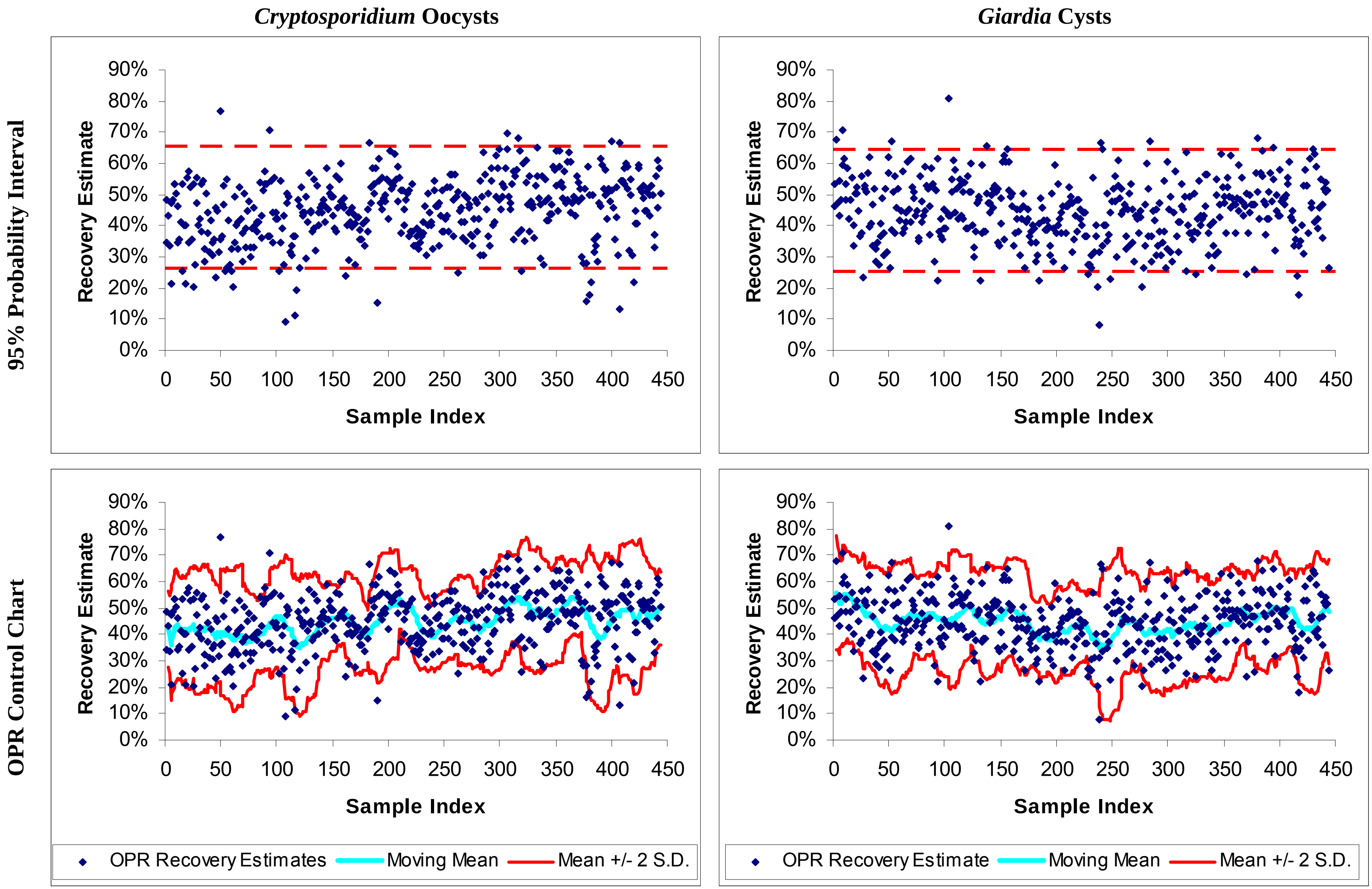


Figure 4.14: Beta-binomial recovery model fitted to OPR dataset from an anonymous Method 1623 approved laboratory



Cryptosporidium Oocysts

Giardia Cysts

Figure 4.15: Temporal variability of OPR recovery estimates

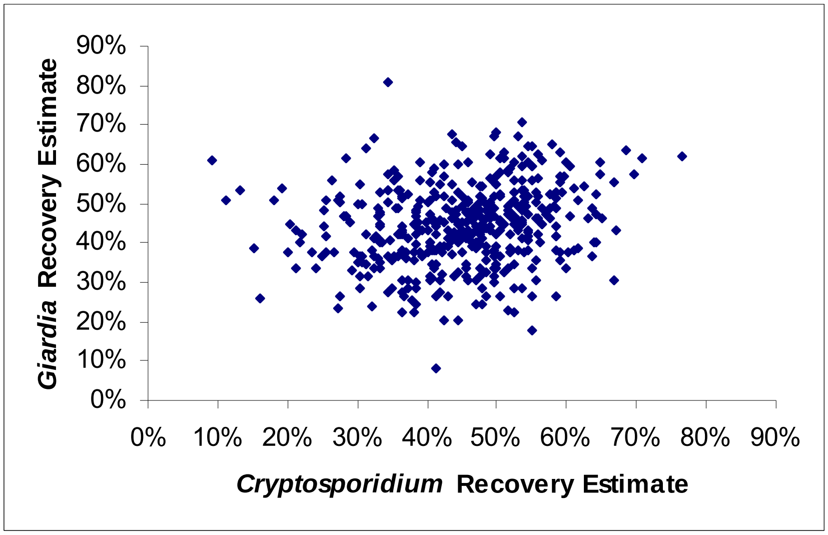


Figure 4.16: Correlation between oocyst and cyst recovery estimates

4.4.3 Quantitative Analysis of Recovery Data Using Bayes' Theorem

It has been demonstrated in Section 4.4.1 that point estimates for the parameters of non-constant analytical recovery distributions must be obtained using a statistical approach that appropriately addresses seeding and analytical errors in the data. Otherwise, biased estimates of the variability in analytical recovery among samples will be obtained because seeding and analytical errors cause recovery estimates to be more variable than analytical recovery itself. Even when an appropriate probabilistic model is used, however, there will be considerable uncertainty in the distribution parameters estimated from small sets of recovery data. Bayes' theorem (Section 3.2.2) is used herein to quantitatively describe uncertainty in unknown parameters using a posterior density function. The posterior density function for the parameters of the beta distribution in the beta-binomial recovery model, for example, represents the relative plausibility of possible parameter values given the model, the available data, and the supplied priors. This section describes the computation of credible regions for the non-constant analytical recovery distribution parameters (*i.e.* contours describing the region in which the true values of the parameters are likely to lie with a specified probability) and the use of Gibbs sampling to generate a Markov chain of the vector of unknown parameters in the model that is representative of the posterior distribution. The former approach obtains approximate posterior densities by numerical integration while the latter is a form of Markov Chain Monte Carlo that is commonly used to evaluate posterior distributions that are too difficult to evaluate by integration.

Computation of approximate credible regions

One of the advantages of Bayesian statistics is that knowledge about uncertainty in the unknown parameters in a model can often be described using the concepts of probability. A 90% credible region for the beta distribution parameters in the beta-binomial recovery model, for example, would be a region of the parameter space (*i.e.* $a > 0$, $b > 0$) within which the analyst is entitled to believe that the true values of the parameters lie (given the model, available data, and chosen priors) with 90%

probability. In one- and two-parameter models, these credible intervals/regions provide a simple graphical display of the uncertainty in the unknown parameters. This section describes a relatively simple approach to compute approximate credible regions using the beta-binomial recovery model. Computation of approximate credible regions using the beta-Poisson or negative binomial recovery models is not discussed herein, but could be done using similar methodology. Schmidt *et al.* (2010a) described an approach to approximate credible regions for the parameters of the beta distribution describing non-constant analytical recovery using he beta-binomial recovery model and relatively uninformative semi-infinite uniform priors (*i.e.* a uniform prior on the parameter space $a > 0$, $b > 0$). The approach of Schmidt *et al.* (2010a) is generalized herein to include other priors represented by the density function $g(a,b)$. It is also demonstrated that the priors used in Schmidt *et al.* (2010a) were inappropriate because they resulted in an improper posterior density function from which probability statements about the unknown parameters in the model (*e.g.* credible regions) cannot be made.

Given the prior $g(a,b)$ for the beta distribution parameters in the beta-binomial recovery model and the likelihood function for these parameters (Equation 4.21), the posterior density function is described by Equation 4.26. The prior can be a distribution (*i.e.* a proper prior) or any other function that is finite and non-negative throughout the parameter space (*i.e.* an improper prior). Gelman *et al.* (2004) defined the posterior density described by a proportionality such as Equation 4.26 as the 'unnormalized posterior density' because the function does not necessarily integrate to unity (in which case the computed densities would not be probability densities). The posterior probability density function is obtained by integration as demonstrated in Equation 4.27. If the integral of the unnormalized density function across the parameter space is not finite, then the posterior is improper (*i.e.* it is not a distribution) and uncertainty in the model parameters cannot be described using probabilities. Accordingly, improper posteriors do not have credible regions unless the posterior is truncated to a region of the parameter space upon which the posterior has a finite integral.

$$f(a,b \mid \{x_i\}) \propto g(a,b) \cdot \left[\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \right)^r \prod_{i=1}^{r} \frac{\Gamma(x_i + a)\Gamma(n_i - x_i + b)}{\Gamma(n_i + a + b)} \right] \tag{4.26}$$

$$f(a,b \mid \{x_i\}) = \frac{g(a,b) \cdot \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \right)^r \prod_{i=1}^{r} \frac{\Gamma(x_i + a)\Gamma(n_i - x_i + b)}{\Gamma(n_i + a + b)}}{\int_0^\infty \int_0^\infty g(a,b) \cdot \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \right)^r \prod_{i=1}^{r} \frac{\Gamma(x_i + a)\Gamma(n_i - x_i + b)}{\Gamma(n_i + a + b)} da \cdot db} \tag{4.27}$$

If the likelihood and prior density associated with specific values of the beta distribution parameters can be calculated accurately, then the following approach can be used to compute approximate credible regions over a rectangular region of the parameter space ($0 < a < a_{max}$, $0 < b < b_{max}$). First, the region was divided into a grid of evenly spaced points and the natural logarithm of the posterior density of each point in the grid was calculated using Equation 4.28 (in which $G(.)$ is the natural logarithm of the gamma function, $\Gamma(.)$). Logarithms were used to avoid overflow errors and small products that are falsely assigned values of zero. The highest computed log-density on the grid was then found and subtracted from each computed log-density (which essentially multiplies all of the density values by a constant). This was also done to avoid overflow errors and to limit false-zeros. The log-densities were then expanded using the exponential function, summed across the grid, and divided by the sum. The normalized density for each point in the grid is an approximate posterior probability for the portion of the grid represented by the point. The probability associated with any region of the parameter space is approximately equal to the sum of the contained normalized densities. A 90% credible region, for example, is any region within which the cumulative probability is 90%. It is most informative, however, to compute a 90% credible region that contains only the most probable points (which is also the smallest 90% credible region). Accordingly, the normalized probabilities were sorted from greatest to least to enable identification of the points in the grid that fall within various approximate credible regions.

$$\begin{aligned} l(a,b) &= \ln(g(a,b)) + r \cdot (G(a+b) - G(a) - G(b)) \\ &+ \sum_{i=1}^{r} (G(x_i + a) + G(n_i - x_i + a) - G(n_i + a + b)) \end{aligned} \tag{4.28}$$

An accurate approximation of the credible regions can be obtained by setting a_{max} and b_{max} to be sufficiently large that the posterior density values along these boundaries are small (*e.g.* so that truncation error is small) and by choosing sufficiently small intervals between grid points to enable good resolution. These various values were determined by trial and error until the resulting credible region plots appeared to be unaffected by further refinements. If the posterior is proper, then the truncation error will become smaller as a_{max} and/or b_{max} are increased and the computed posterior probabilities will converge upon the actual values as the resolution is increased. If the posterior is improper, however, then changing a_{max} and/or b_{max} and/or the resolution of the grid will affect the computed probabilities and credible regions (although the effect may be indiscernible). The algorithm is analogous to a numerical integration and the associated sum will never truly converge unless the posterior is proper. Even if the posterior density function is proper, poorly approximated credible regions will result if posterior densities outside of the grid are relatively high because the computed credible regions may intersect the edge of the grid and will be incorrect. Additionally, if the grid is too coarse (*i.e.* it is divided up into too few evenly spaced points), then the poor resolution will result in imprecise credible regions.

Approximate credible regions were computed for the beta distribution parameters estimated from the Table 4.4 recovery data using the beta-binomial recovery model and the semi-infinite uniform priors that were used in Schmidt *et al.* (2010a). These analyses were completed in Microsoft Excel^{TM} using code for the algorithm described above that was written in the Visual Basic Editor. The relationship $b_{max} = a_{max} \cdot (1/\mu_p - 1)$, which is based upon the estimated mean analytical recovery μ_p, was used to optimize efficiency because the highest posterior densities are expected along or near this line. Figure 4.17a is an example in which the credible regions are inaccurate because a_{max} and b_{max} are too low and the grid spacing is large (which results in poor resolution). Figure 4.17b shows approximate credible regions associated with a larger grid (with better resolution) as well as relevant

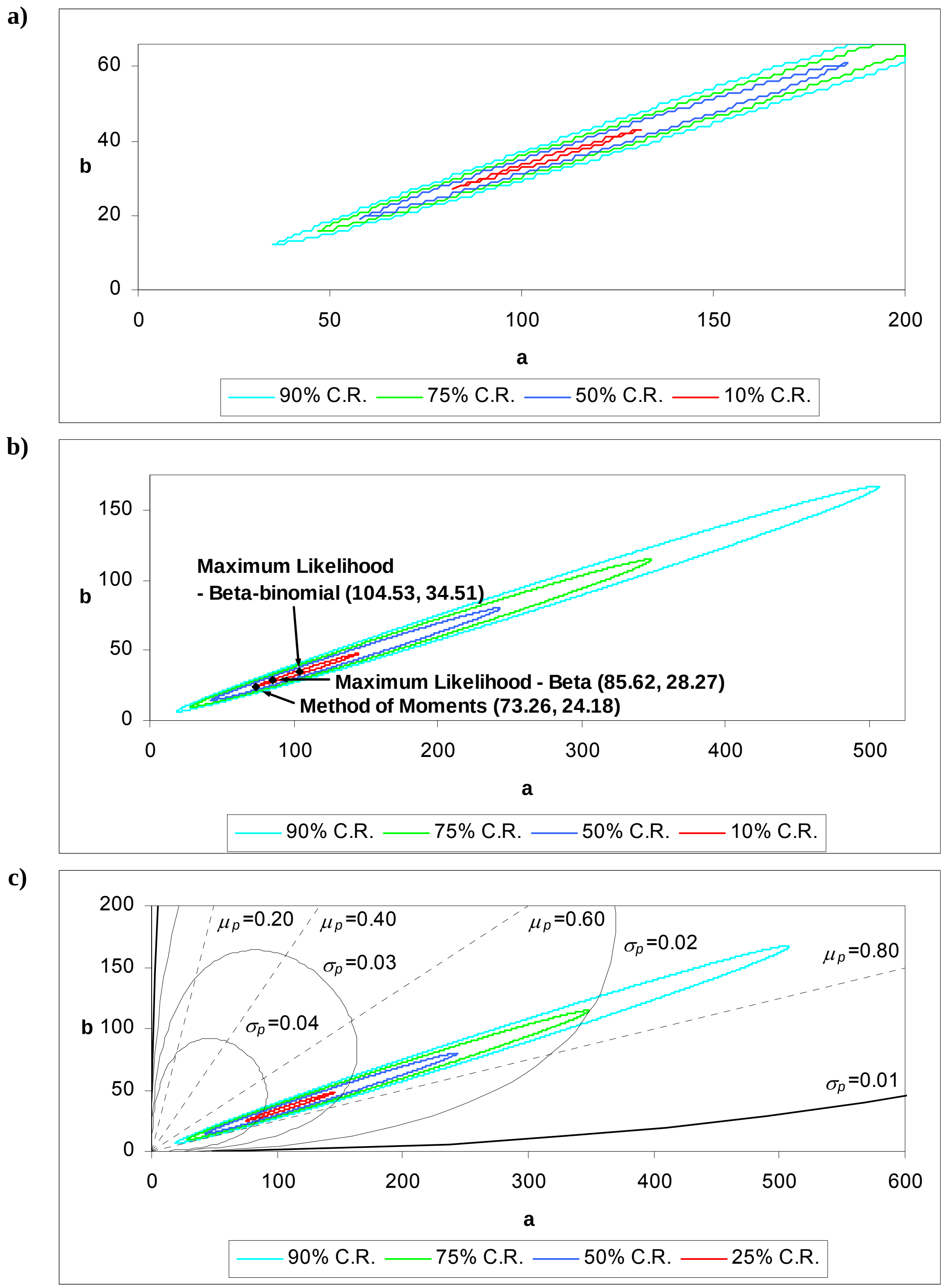


Figure 4.17: Credible regions calculated from the Table 4.4 data

a) inaccurate credible regions that extend beyond the selected parameter space and have poor resolution, b) approximate credible regions plotted with point estimates from various parameter estimation approaches (Schmidt *et al.*, 2010a), c) approximate credible regions plotted with contours for various means and standard deviations of analytical recovery.

point estimates of the beta distribution parameters from Table 4.5. This figure shows that there is considerable uncertainty in the estimated parameters of the beta distribution that describes non-constant analytical recovery. The slope of the long axis of the credible regions relates to the mean of the beta distribution because $b = a \cdot (1/\mu_p - 1)$. The proximity of a pair of beta parameter values to the origin relates to the variance because low values of a,b correspond to higher variance. The relationships between the beta distribution parameters and the associated mean and standard deviation of analytical recovery are illustrated in Figure 4.17c. This figure shows that a small reduction in the standard deviation of the analytical recovery can result in a substantial increase in the beta distribution parameter values. This approach for computing approximate credible regions may not work if the variability in analytical recovery among samples is very low because impractically large values of a_{max} and b_{max} will be needed and the number of grid points required to ensure good resolution would be prohibitively large.

The presentation of credible regions associated with a posterior density function is based upon an assumption that the posterior is proper. If a posterior is improper (*i.e.* its integral over the parameter space is not finite), then it is impossible to describe uncertainty in the unknown parameters in the model using probability concepts such as credible regions. The uniform improper priors used in Schmidt *et al.* (2010a) and in the analyses above do yield a posterior density function that quantitatively describes uncertainty in the beta distribution parameters (*e.g.* relative posterior densities for various parameter pairs can be compared and the shape of the posterior can be displayed by contours), but the posterior is improper and credible regions do not exist. The beta distribution in this model describes the variation in recovery among samples and there is always some possibility that recovery is constant and that all of the observed variability is due to analytical error (and seeding error where applicable). If analytical recovery is a constant, then the beta-binomial recovery model simplifies to the 'binomial-only' recovery model (Section 4.4.1). Accordingly, as a and b increase along any line $b = a \cdot (1/\mu_p - 1)$, the likelihood can never fall below the 'binomial-only' likelihood associated with μ_p. The posterior obtained using semi-infinite uniform priors is improper because the likelihood function does not converge asymptotically upon zero as the beta distribution parameters increase. Similarly, the likelihoods associated with the beta-Poisson and negative binomial recovery models will converge upon the 'Poisson-only' recovery model as a and b approach infinity or as α approaches infinity and β approaches zero. The approximate credible regions plotted in Figure 4.15 appeared to converge because the 'binomial-only' likelihoods were very small and the effects of further refinement were indiscernible. The presented credible regions do correctly portray the shape of the posterior density function with contours, but probability statements cannot be made about the parameters falling within any contour (unless the proper uniform prior on the region $0 < a < a_{max}$, $0 < b < b_{max}$ is used). Computation of credible regions using other priors is not discussed herein.

Evaluating posterior distributions using Gibbs sampling

Information about the posterior distribution of the non-constant analytical recovery distribution parameters (a,b or α,β) can also be obtained by Gibbs sampling if the posterior is proper. Development of the Gibbs sampling algorithm for the beta-binomial recovery model is discussed herein. Information to aid development of Gibbs sampling algorithms for the beta-Poisson and negative binomial recovery models is included in Appendix B.1. As described in Section 3.2.2, Gibbs sampling can be used to generate a Markov chain of values that are collectively representative of a posterior distribution (assuming convergence). This is done by supplying initial values for all unknown parameters in the model and then repeating an iterative loop in which each unknown parameter is replaced with a random value from its associated conditional posterior distribution. The steps involved in setting up a Gibbs sampling algorithm include (1) application of Bayes' theorem (with specified priors) to obtain a joint posterior density function for all of the unknown parameters in the model, (2) derivation of the conditional posterior density function for each unknown parameter in the model, (3) preparation of algorithms to draw random numbers from each of the conditional posterior distributions, and (4) specification of initial values for each of the unknown parameters. Hobert and Casella (1996) argue that determination of the propriety of the posterior (*i.e.* whether or not the posterior density has a finite integral across the parameter space) must precede these steps because Gibbs sampling associated with an improper posterior may work without any obvious 'red flags' even though the results do not converge (*i.e.* the results will never truly represent the posterior).

Using $g(a,b)$ to represent the prior for the beta distribution parameters, application of Bayes' theorem to the beta-binomial model (Equation 4.3) yields the joint posterior density function represented by Equation 4.29. In this equation, $\{x_i\}$ are the counts of observed particles and $\{n_i\}$ are parameters with known values (*i.e.* numbers of seeded particles). The beta distribution parameters (a,b) are the parameters of interest, while the actual analytical recovery values associated with the samples in the recovery dataset ($\{p_i\}$) are nuisance parameters (for which it is not necessary to specify a prior). It is possible to integrate the nuisance parameters out of the model (as in Equation 4.4 or Equation 4.26), but this results in a more complicated Gibbs sampling algorithm (results not shown).

$$f(a,b,\{p_i\}|\{x_i\}) \propto g(a,b)\cdot\prod_{i=1}^{r}\left[\frac{n_i!}{x_i!(n_i-x_i)!}p_i^{x_i}(1-p_i)^{n_i-x_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p_i^{a-1}(1-p_i)^{b-1}\right] \quad (4.29)$$

The conditional posterior distribution functions for a, b, and p_i (as derived from Equation 4.29) are summarized in Equations 4.30, 4.31, and 4.32 respectively. In each equation, all terms that do not include the unknown parameter are constants (because all the other parameters have specified values in the conditional posterior distribution function) and can be dropped from proportionalities. The equations have been simplified algebraically where possible (*e.g.* by expanding the product in Equation 4.29 and by dropping any constant terms from the proportionality).

$$Dfc(a) \propto g(a\,|\,b)\cdot\left(\frac{\Gamma(a+b)}{\Gamma(a)}\right)^{r}\left(\prod_{i=1}^{r}p_i\right)^{a} \quad (4.30)$$

$$Dfc(b) \propto g(b\,|\,a)\cdot\left(\frac{\Gamma(a+b)}{\Gamma(b)}\right)^{r}\left(\prod_{i=1}^{r}(1-p_i)\right)^{b} \quad (4.31)$$

$$Dfc(p_i) \propto p_i^{x_i+a-1}(1-p_i)^{n_i-x_i+b-1} \rightarrow p_i \sim BETA(x_i+a, n_i-x_i+b) \quad (4.32)$$

As shown in Equation 4.32, the conditional posterior distribution of p_i is a beta distribution. Accordingly, random values of p_i can be generated using widely available algorithms that generate beta-distributed random numbers. The conditional posterior distributions for the beta distribution parameters, however, cannot be arranged in the form of density functions for which random number generation algorithms are readily available. A simple acceptance-rejection algorithm was developed to sample from the conditional posterior distributions of a and b. These conditional posterior distribution functions both have a form that can be represented by Equation 4.33 (in which $g(\theta)$ is the conditional prior for the parameter θ given the other beta distribution parameter). The first step of the utilized acceptance-rejection algorithm is to determine the maximum value of the distribution function (f_{max}). A greedy optimization algorithm (a simple local optimization algorithm) can be used if the distribution is unimodal because it will reliably converge upon the global maximum. Proof that these conditional posterior distributions are unimodal given the priors considered herein is included in Appendix A. Next, boundaries of θ (θ_{min}, θ_{max}) are found outside of which the conditional posterior density was small (*e.g.* $< 10^{-10} \cdot f_{max}$). A uniform distribution on this interval is then used to generate a test value (θ_{test}) which is accepted if a second generated uniform value (f_{test}) on the interval $(0, f_{max})$ is less than $f(\theta_{test})$. Figure 4.18 illustrates this acceptance rejection algorithm and the results of two iterations (of which the first is rejected and the second is accepted). This algorithm is approximate because the interval (θ_{min}, θ_{max}) truncates the tails of the conditional posterior distribution. This algorithm can be inefficient if the distribution is wide (because many iterations may be needed before a generated value is accepted), but it is relatively simple. Development of other acceptance-rejection algorithms that may be more efficient and that may avoid truncation error are not discussed herein.

$$f(\theta) \propto g(\theta) \cdot \left(\frac{\Gamma(\theta + s)}{\Gamma(\theta)} \right)^r t^{\theta} \tag{4.33}$$

The final requirement to set up the Gibbs sampling algorithm is to specify an initial value for each of the parameters. Any value within the feasible parameter space (*i.e.* $a > 0$, $b > 0$, $0 < p_i < 1$) is acceptable, although a shorter burn-in (number of iterations that are discarded so that the retained portion of the Markov Chain is minimally impacted by the supplied initial parameter values) is typically possible if relatively probable parameter values are chosen. Method of moments or maximum likelihood point estimates can be used to initialize the beta distribution parameters (a,b). No initial values are required for $\{p_i\}$ if the Gibbs sampling algorithm is set up to generate $\{p_i\}$ first. The Gibbs sampling algorithm is illustrated in the form of a flowchart in Figure 4.19.

The Gibbs sampling algorithm will generate a Markov Chain of the vector of parameter values $(\{p_i\},a,b)$ that represents the joint posterior distribution (Equation 4.29) assuming appropriate convergence. Good convergence requires that Gibbs sampling be run long enough (following an appropriate burn-in) that the relative frequency with which a parameter falls in a region of parameter space is approximately equal to the associated posterior probability. A Markov chain with good mixing (*i.e.* one that does not linger in a small region of the parameter space for a large number of

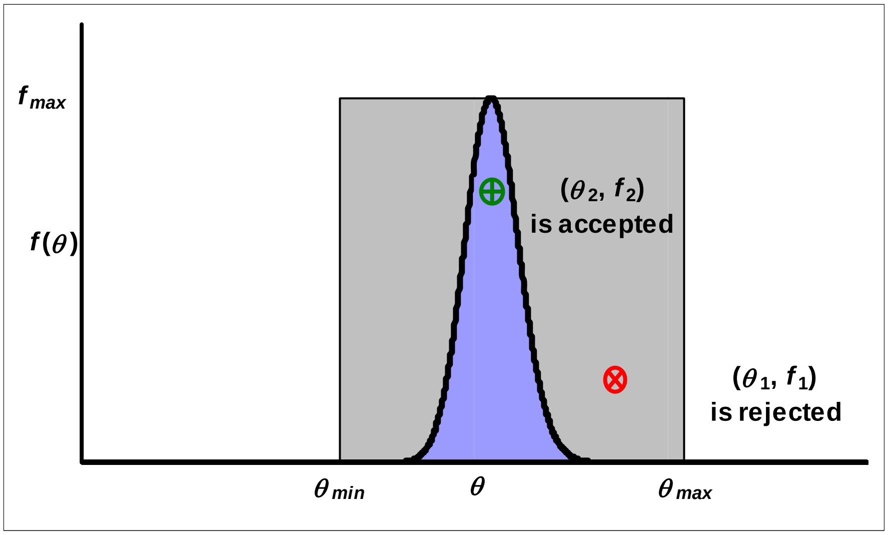


Figure 4.18: Acceptance-rejection sampling algorithm

The two uniform random deviates (θ_i, f_i) represent a point within the pale grey box. If $f_i \leq f(\theta_i)$ (*i.e.* the pair of deviates represents a point under the curve), then θ_i is accepted as a random deviate from the distribution of θ.

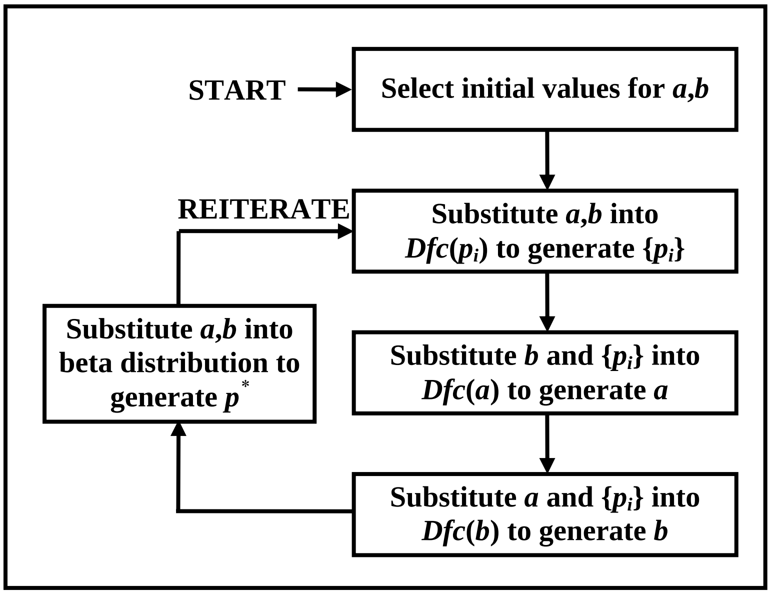

START
Select initial values for *a,b*
REITERATE
Substitute *a,b* into $Dfc(p_i)$ to generate $\{p_i\}$
Substitute *b* and $\{p_i\}$ into $Dfc(a)$ to generate *a*
Substitute *a* and $\{p_i\}$ into $Dfc(b)$ to generate *b*
Substitute *a,b* into beta distribution to generate p^*

Figure 4.19: Gibbs sampling algorithm for the beta-binomial recovery model

The Gibbs sampling algorithm begins with a burn-in and ends after a user-specified number of iterations. If a single value of p^* is generated in each iteration after the burn-in and the Markov chain has converged, then the set of p^* values will represent the posterior predictive distribution of analytical recovery.

iterations) will require a shorter burn-in and a smaller number of iterations to reach satisfactory convergence. Good convergence can be demonstrated by repeatedly running the Gibbs sampling algorithm for a specified number of iterations (particularly if each run uses different initial values of the unknown parameters from throughout the parameter space) and obtaining very similar estimated posterior distributions and probabilities. If very different results are obtained, then either more iterations are required (possibly including a longer burn-in) or the Markov chain does not converge.

The generated values of $\{p_i\}$ in this Gibbs sampling application are generally not useful unless the exact values of analytical recovery in the recovery experiment samples are of interest. The generated pairs of beta distribution parameter estimates (a,b), however, can be used to answer many practical questions. For example, each pair (after the burn-in) can be transformed to its associated mean or standard deviation to represent the posterior distribution for the mean or standard deviation of analytical recovery. The posterior distribution of the standard deviation, for example, would describe what the analyst is entitled to believe about the actual standard deviation of analytical recovery among samples given the model (which accounts for analytical error), available recovery data, and the chosen priors. A posterior predictive distribution of analytical recovery can also be generated to describe what the analyst is entitled to believe about the distribution of possible future values of analytical recovery given the beta distribution model for variability in recovery and uncertainty in its parameters. This is accomplished by generating a single beta-distributed deviate (p^*) from each pair of beta distribution parameters in the Markov chain (after the burn-in) as shown in Figure 4.19.

Analysis of Table 4.4 data using Gibbs sampling and uniform priors

The Gibbs sampling algorithm presented above was used to analyze the Table 4.4 recovery data (assuming that a beta-binomial recovery model is appropriate for the data). The analyses were completed using the Visual Basic Editor in Microsoft Excel^{TM}. Relatively uninformative improper uniform priors (on the parameter space $a > 0$, $b > 0$) were used. The results of a run with 31,000 iterations (of which the first 1000 are discarded as burn-in) are plotted in Figure 4.20. The top left panel plots the sequence of beta distribution parameter values, the top right panel plots the sequence of analytical recovery values that is assumed to represent the posterior predictive distribution of analytical recovery, and the bottom panels show the sequences of values that are assumed to represent the posterior distributions of the mean and standard deviation of analytical recovery. These results demonstrate somewhat poor mixing of the beta distribution parameters and standard deviation of analytical recovery, but good mixing of the simulated recovery values and the mean recovery. Three repeated runs yielded very similar empirical cumulative distribution functions (coinciding curves with minimal noise) from the sequences of recovery values that were assumed to represent the posterior predictive distribution of analytical recovery (results not shown). Such results can easily be mistaken to indicate good convergence, which is disproven below. As demonstrated above in the discussion of credible regions, the posterior of the beta-binomial recovery model with semi-infinite uniform priors is improper. The beta distribution parameters have the opportunity to diverge to infinity without returning to low values because the posterior density does not converge asymptotically upon zero as

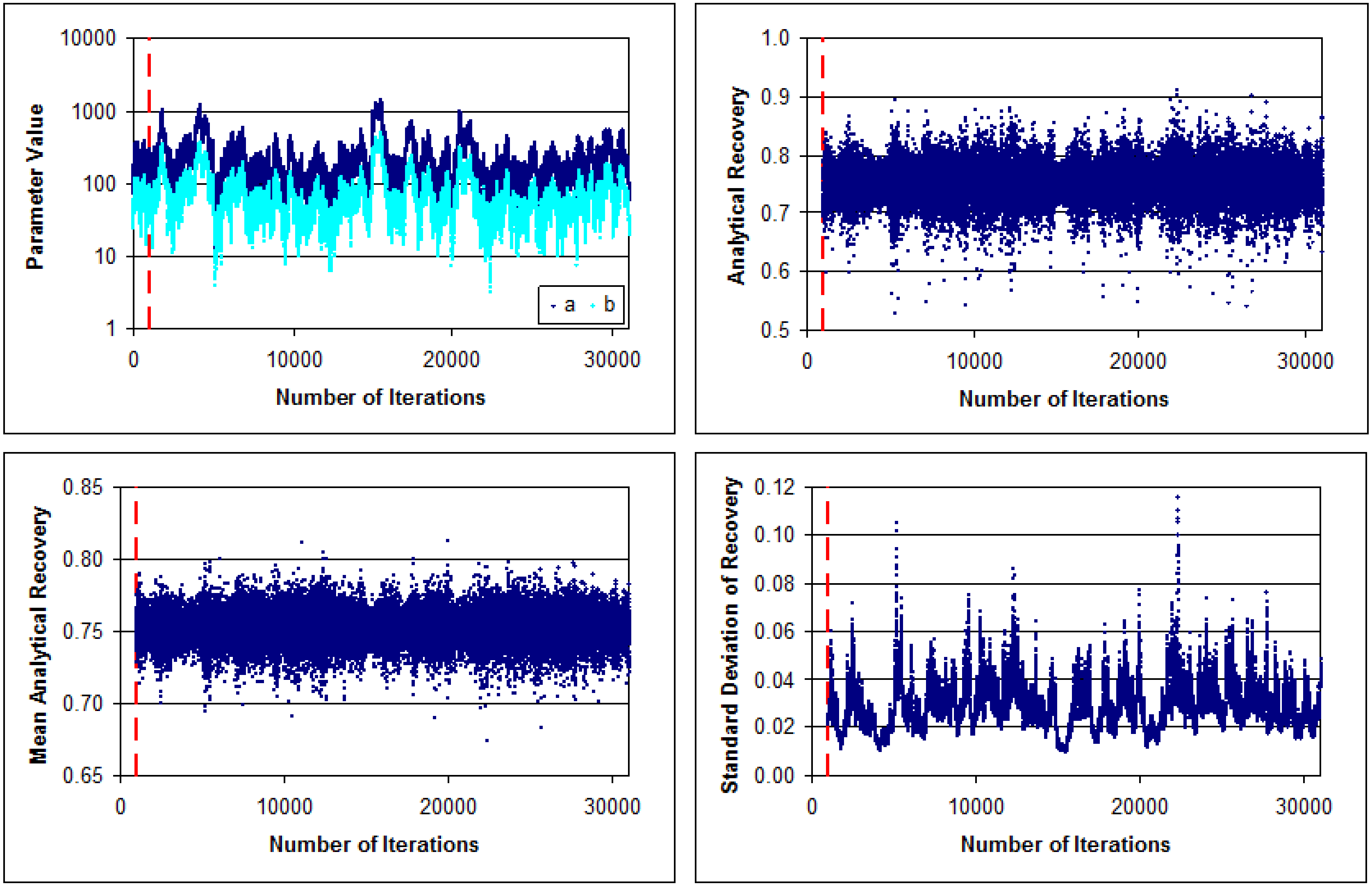


Figure 4.20: Sequences of values obtained using Gibbs sampling with the Table 4.4 data and the beta-binomial model

These results are based on 31,000 iterations, with a 1000-iteration burn-in, and use improper uniform priors for the beta distribution parameters.

these parameters increase. The above Gibbs sampling example, therefore, only has an apparent convergence because the parameters stayed near the mode of the posterior without (yet) diverging to very different outcomes. Possible evidence of poor convergence may include long Gibbs sampling runs that suddenly diverge to unusual parameter values (without ever returning to typical values) or Gibbs sampling runs with diverse initial parameter values that have vastly different results.

Figure 4.21 shows the results of three additional Gibbs sampling runs to illustrate convergence issues associated with uniform priors. Figures 4.21a and 4.21b use semi-infinite uniform priors for the beta distribution parameters and different initial parameter values. Figure 4.21a starts with low parameter values ($a = 1$, $b = 1$) and the Markov chain quickly reaches the typical values obtained in Figure 4.20. If a similar outcome resulted from all initial parameter values, then a small burn-in would be acceptable and it would suggest that the Gibbs sampling process has converged if repeated runs yield similar results (after burn-in). Figure 4.21b starts with higher parameter values ($a = 10000$, $b = 10000$) and the resulting Gibbs sampling sequence is very different. This confirms that the sequence of values in a Markov chain of only 30,000 iterations (following a burn-in of 1000 iterations) depends upon the supplied initial parameter values and has therefore not (yet) converged despite the apparent convergence demonstrated by repeated runs with more typical initial parameter values. If the Gibbs sampling sequence were continued for many more iterations, then the parameter values would likely diverge to infinity rather than return to the typical parameter values obtained in the other runs because the posterior associated with improper uniform priors is improper (as discussed above). Figure 4.21c uses the same initial parameter values as Figure 4.21b but uses the proper uniform prior ($0 < a < 10000$, $0 < b < 10000$). In this example, the sequence of parameter values falls slowly to the typical values shown in Figures 4.20 and 4.21a (indicating the need for a longer burn-in than 1000 iterations). The posterior is proper because the prior is proper, and so this Markov chain will eventually converge.

Millions of iterations (or more) may be required to sample the entire parameter space and yield a truly converged representation of the posterior distribution (particularly if mixing is poor). In the present example (with the proper uniform prior), a Gibbs sampling run starting with maximum likelihood or method of moments parameter estimates will reasonably characterize the posterior distribution after only 30,000 iterations (and a burn-in of 1000 iterations), but the results are not truly converged because a small fraction of such runs will sample the higher beta parameter values and yield very different results. Additionally, these repeated runs have some noise in the estimated posterior probabilities (*i.e.* random variation between relative frequencies and actual posterior probabilities) which is especially prevalent in the tails of the posterior. Although it is convenient to assure a proper posterior by using proper priors, a proper uniform prior should only be used to represent known boundaries of the parameters (in the absence of other knowledge about the parameters) rather than to restrict the parameter space arbitrarily. The proper uniform priors used in Figure 4.21c, for example, reject any possibility of higher parameter values. As more data are collected, the resulting posterior distributions will gradually converge upon the true values of the unknown parameters with diminishing variability (*i.e.* less uncertainty in parameter estimates) provided that the prior does not exclude the true parameter values.

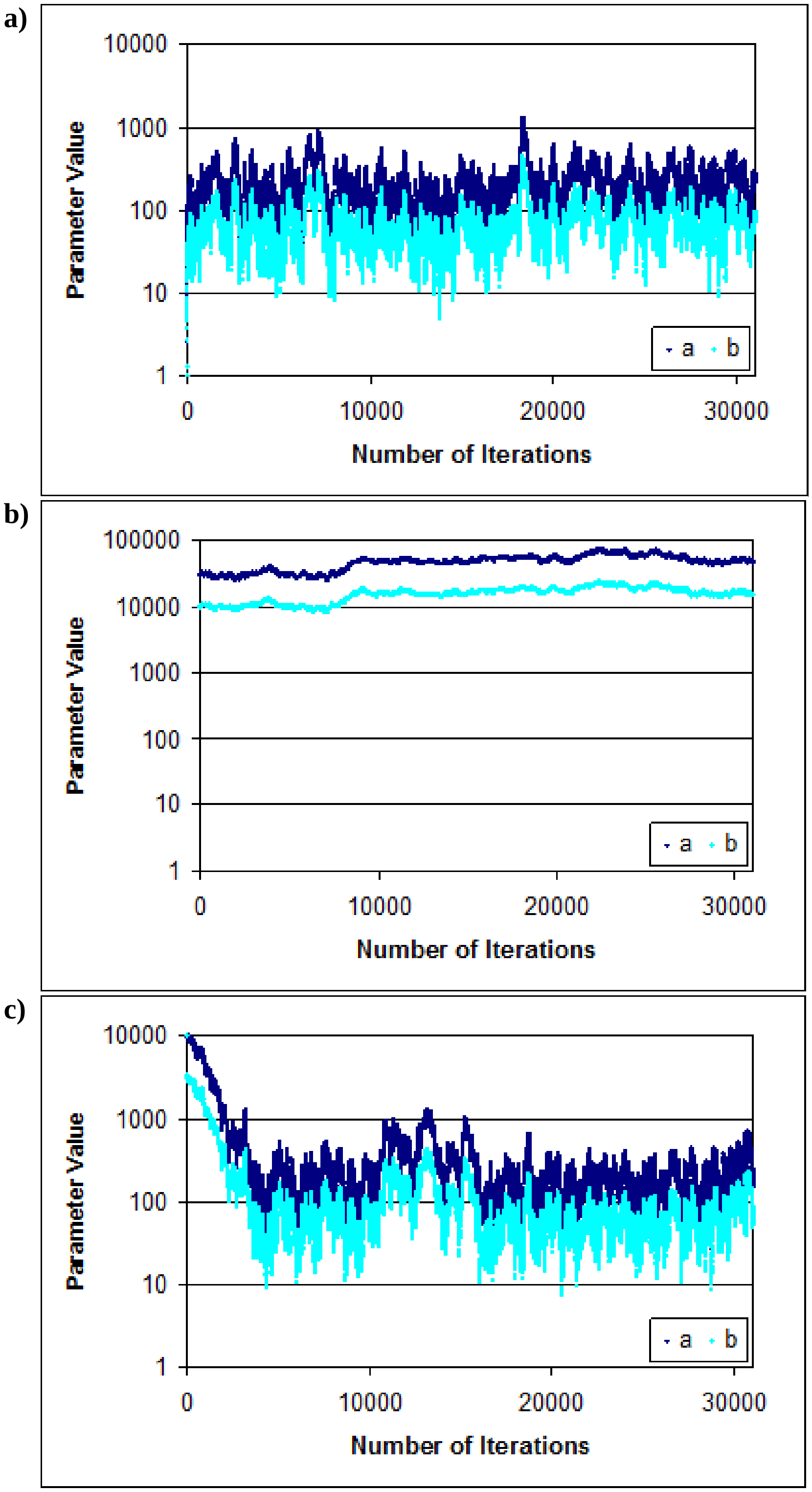


Figure 4.21: Example Gibbs sampling sequences used to evaluate convergence

These Gibbs sampling results are based on analysis of the Table 4.4 recovery data using a beta-binomial recovery model with uniform priors. The priors and initial beta distribution parameters for each panel are (a) improper semi-infinite uniform priors with a_0 = 1, b_0 = 1, (b) improper semi-infinite uniform priors with a_0 = 10000, b_0 = 10000, and (c) proper uniform priors on the parameter space 0 < a < 10000, 0 < b < 10000 with a_0 = 10000, b_0 = 10000.

Analysis of Table 4.4 data using Gibbs sampling and several relatively uninformative priors

An important feature of Bayes' theorem is that it enables integration of subjective information about the unknown parameters in a model (the prior) with information provided by the available data (the likelihood function). The analyst should use a prior that represents whatever reasonably grounded subjective information can be provided (*e.g.* from preliminary or similar experiments). The resulting posterior is partially subjective because it depends upon the analyst's choice of prior, and an inappropriate prior can bias the results. Weakly informative priors yield relatively objective results because the posterior is most heavily influenced by the likelihood function (which is based only upon the model and the available data). Relatively uninformative priors are considered herein, but the presented methods can relatively easily be adapted to include informative priors. An uninformative prior does not restrict the parameter space of the unknown parameters and is often relatively flat so that the resulting posterior is much more heavily influenced by the likelihood function than the prior. There is, however, no such thing as a truly uninformative prior because any prior will affect the posterior distribution of the unknown parameters (or transformations thereof). For example, in the beta-binomial recovery model, the uniform improper prior on the parameter space $a > 0$, $b > 0$ is relatively uninformative because it indicates that all possible pairs of beta distribution parameters in the parameter space are equally likely. The resulting posterior density function has identical shape to the likelihood function because the prior is uniform and all posterior information about uncertainty in the beta distribution parameters can accordingly be attributed to the likelihood function. This prior is not truly uninformative, however, because it favours low standard deviations of analytical recovery: smaller standard deviations represent a larger portion of the a,b parameter space (as shown in Figure 4.15c) and will accordingly have larger prior densities than smaller standard deviations. Several alternative relatively uninformative priors are considered herein. The purpose is not to select a universally applicable uninformative prior for the problem at hand, but to illustrate the effects of various priors upon the posterior and associated Gibbs sampling results.

Two families of priors are considered herein for the parameter space $a > 0$, $b > 0$. The first has a general formula $g(a,b) = (ab)^{-s}$ and is improper. Two alternatives from this family of priors are considered with $s = 0$ and $s = 1$. The former is the improper uniform prior used above, which has been demonstrated to yield an improper posterior. The latter is based on a recommendation by Jeffreys (1961) that the uninformative prior $1/\theta$ should be used for a continuous parameter with semi-infinite domain ($\theta > 0$). Improper priors may lead to improper posteriors, which would preclude the use of Markov Chain Monte Carlo methods such as Gibbs sampling to evaluate the posterior. Accordingly, relatively uninformative proper priors should also be considered because these will yield proper posteriors. Petterson *et al.* (2007) used a proper prior in which $\log_{10}(a)$ and $\log_{10}(b)$ were uniform on the interval (-10,10). This is analogous to the prior $g(a,b) = 1/ab$ with the restricted parameter space ($10^{-10} < a < 10^{10}$, $10^{-10} < b < 10^{10}$). This particular prior is not considered herein. The second considered family is the proper exponential priors $g(a,b) = exp(-(a + b) / t)$, which are relatively uninformative for large values of t. In the present example, $t = 500$ is used, for which the prior probability that $a + b < 1500$ is approximately 95%.

The Gibbs sampling algorithm presented above is used in each of these examples to generate a sequence of 30,000 iterations (following a burn-in of 1000 iterations). It is acknowledged that convergence may be poor after so few iterations and that convergence may never exist if the prior is improper. This number of iterations was used for quick and direct comparison of the results, not a precise evaluation of the posteriors. The empirical cumulative distribution function of the sequence of analytical recovery values that are assumed to represent the posterior predictive distribution of analytical recovery is plotted for each of the considered priors in Figure 4.22. The non-constant analytical recovery distribution associated with the maximum likelihood estimates (a = 104.53, b = 34.51) is also plotted for comparison.

The similarity among the various curves plotted in Figure 4.22 suggests that the priors are relatively uninformative (because the curves are not substantially different) and that accounting for uncertainty in the beta distribution parameters does not substantially improve knowledge about the variability in analytical recovery relative to using point estimates obtained by maximum likelihood. These conclusions are, however, based on an assumption that the first three curves represent converged posterior predictive distributions. It has been proven that the uniform prior yields an improper posterior and that Gibbs sampling will not converge with this prior. It is unknown whether the prior $g(a,b) = (ab)^{-1}$ will yield a proper posterior, and determination of propriety is beyond the scope of the present investigation. The exponential prior will yield a proper posterior, but it is not

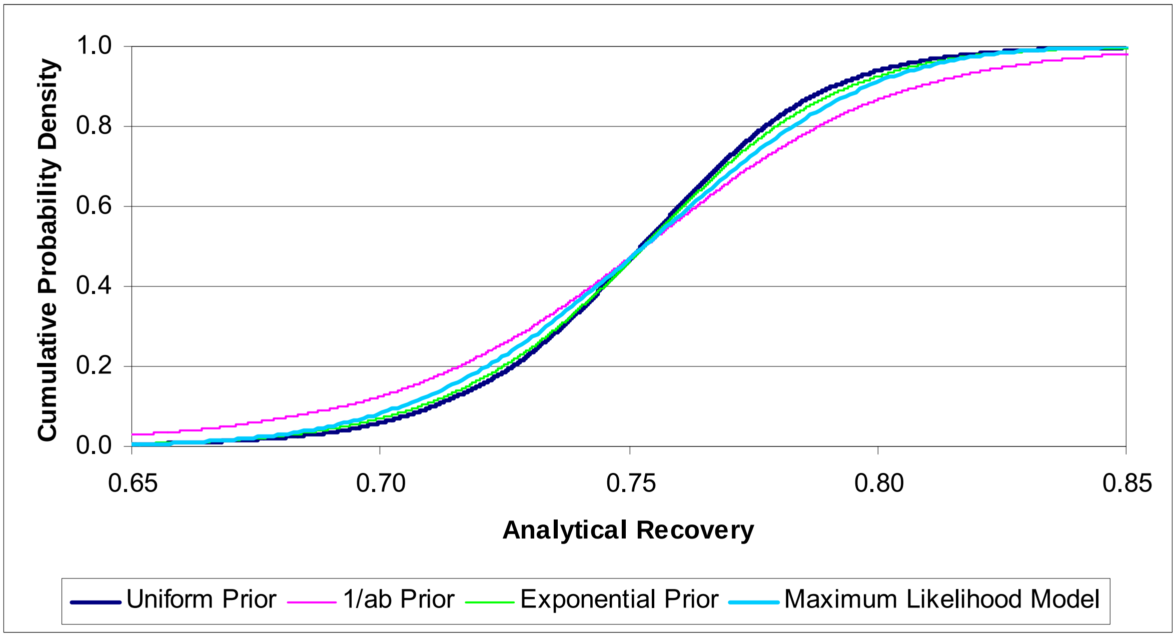


Figure 4.22: Alternative distributions modelling variability in analytical recovery

The first three curves represent the empirical cumulative distribution function of the analytical recovery values generated by Gibbs sampling that are assumed to represent the posterior predictive distribution. The three alternative priors were the improper uniform prior (across the parameter space $a > 0$, $b > 0$), the Jeffreys prior $g(a,b) = 1/ab$, and the exponential prior $g(a,b) = exp(-(a + b) / 500)$. The maximum likelihood model is the beta distribution with parameters equal to the beta-binomial maximum likelihood estimates.

known to what extent the estimated posterior predictive distribution has converged after only 30,000 iterations. Accordingly, these three curves represent sequences of values near the mode of the posterior but may not represent posterior predictive distributions. Nonetheless, these three curves do show some slight differences that are anticipated from the effects of the various priors upon the beta distribution parameters. The uniform prior, for example, will favour low values of the standard deviation of analytical recovery and the associated curve is the steepest (indicating less variable analytical recovery). The prior $g(a,b) = (ab)^{-1}$ favours low beta distribution parameter values, which correspond to higher values of the standard deviation of analytical recovery, and the associated curve is the least steep of the alternatives. The exponential prior also slightly favours low beta distribution parameter values, but the selected prior is wide and uninformative (*i.e.* with a relatively high value of t) and the associated curve is close to the curve associated with the uniform prior. The model for non-constant analytical recovery associated with the beta-binomial maximum likelihood estimates falls between these posterior predictive distributions for recovery.

It is always important in Bayesian analyses to use appropriate probabilistic models and priors. If the model does not appropriately represent the random errors (*e.g.* seeding error, analytical error, and non-constant analytical recovery) in the data, then the resulting posterior distributions will be incorrect. Relatively uninformative priors should always be used unless there is sound reasoning (*e.g.* preliminary data or results from similar experiments) to support use of an informative prior. Even if a relatively uninformative prior is used, the choice of prior can affect the resulting posterior distribution (*e.g.* Figure 4.22). In general, the prior will have greatest effect upon the posterior when the likelihood function contributes little information about the parameters (*e.g.* when recovery data correspond to imprecise recovery estimates because of substantial seeding and analytical errors). Accordingly, the results of Bayesian analyses can be quite subjective when very imprecise data (or few data) are used.

4.4.4 Case Study: Analytical Recovery of *Schistosoma* Eggs Enumerated in Feces

A new method to enumerate *Schistosoma mansoni* eggs in human feces, which enables enumeration of larger fecal samples than conventional methods, has been developed to enable better detection of low egg burdens (Teixeira *et al.*, 2007). The method is susceptible to many losses, however, because many steps are required to purify the sediment and to separate the eggs from the purified sediment. The conventional thick smear technique (Katz *et al.*, 1972) enables enumeration of only very small feces samples and is only sensitive (*i.e.* likely to yield positive diagnosis) when infected individuals have high egg burdens. To evaluate the sensitivity of the method, a recovery experiment was conducted with multiple 30g feces samples that were seeded with known quantities of eggs. Seeding errors (*e.g.* as discussed in Section 4.1.2) were assumed to be minimal in this experiment. The results are summarized in Table 4.6.

Table 4.6: Analytical recovery of *Schistosoma mansoni* eggs in human feces

(Teixeira *et al.*, 2007)

	Number of Seeded Eggs in 30g Feces						
	3	**7**	**10**	**20**	**30**	**40**	**60**
Number of Observed Eggs	0	1	1	3	3	4	5
	0	1	0	1	3	2	3
	0	2	0	0	4	7	6
	1	0	0	0	1	1	2
	0	0	0	0	0	3	2
	0	3	0	2	5	10	3
	1	0	2	0	7	5	
	0	0	2	3	3	6	
	0	0	3	1	0	5	
	0	3	2	1	4	5	
Avg. % Recovery	6.7%	14.3%	10.0%	5.5%	10.0%	12.0%	5.8%
% Sensitivity	20%	50%	50%	60%	80%	100%	100%

The beta-binomial recovery model is most appropriate to describe the random errors in this experiment because the numbers of seeded eggs were assumed to be precisely known. If there were any losses during seeding, the estimated mean analytical recovery would be biased low and the variability of analytical recovery would be biased high (both of which are conservative biases when estimating egg burdens, quantifying uncertainty in estimated egg burdens, or quantifying the sensitivity of the enumeration method). Assuming that the number of seeded eggs did not impact the analytical recovery of the enumeration method, all the data in Table 4.6 can be pooled. The resulting maximum likelihood estimates for the parameters of the beta distribution that describes variability in analytical recovery among samples are $a = 3.34$, $b = 32.90$. The maximum likelihood estimate of the constant analytical recovery in the 'binomial-only' recovery model is approximately 0.0904 (which can be determined explicitly as the sum of the observed eggs divided by the sum of the seeded eggs). The log-likelihoods associated with these two models are -108.22 and -112.17. Therefore, expanding the 'binomial-only' recovery model to include non-constant analytical recovery (and an extra fitted parameter) yielded a maximum likelihood that is only 51.7 times greater. This is not compelling evidence for non-constant analytical recovery because the beta-binomial likelihoods will not become trivially small as the standard deviation of analytical recovery falls to zero (*i.e.* as the beta distribution parameters approach infinity for any specified mean recovery μ_p along the line $b = a \cdot (1/\mu_p - 1)$). Figure 4.23 shows the posterior density associated with improper semi-infinite uniform priors for various values of the beta distribution parameter a with the mean analytical recovery set to 9.04%. Gibbs sampling cannot be used to evaluate the posterior associated with these uniform priors because the posterior is improper.

Four priors were considered herein for the analysis of this recovery dataset: the improper priors $g(a,b) = (ab)^{-0.5}$ and $g(a,b) = (ab)^{-1}$, and the proper exponential priors $g(a,b) = exp(-(a + b) / 100)$, and

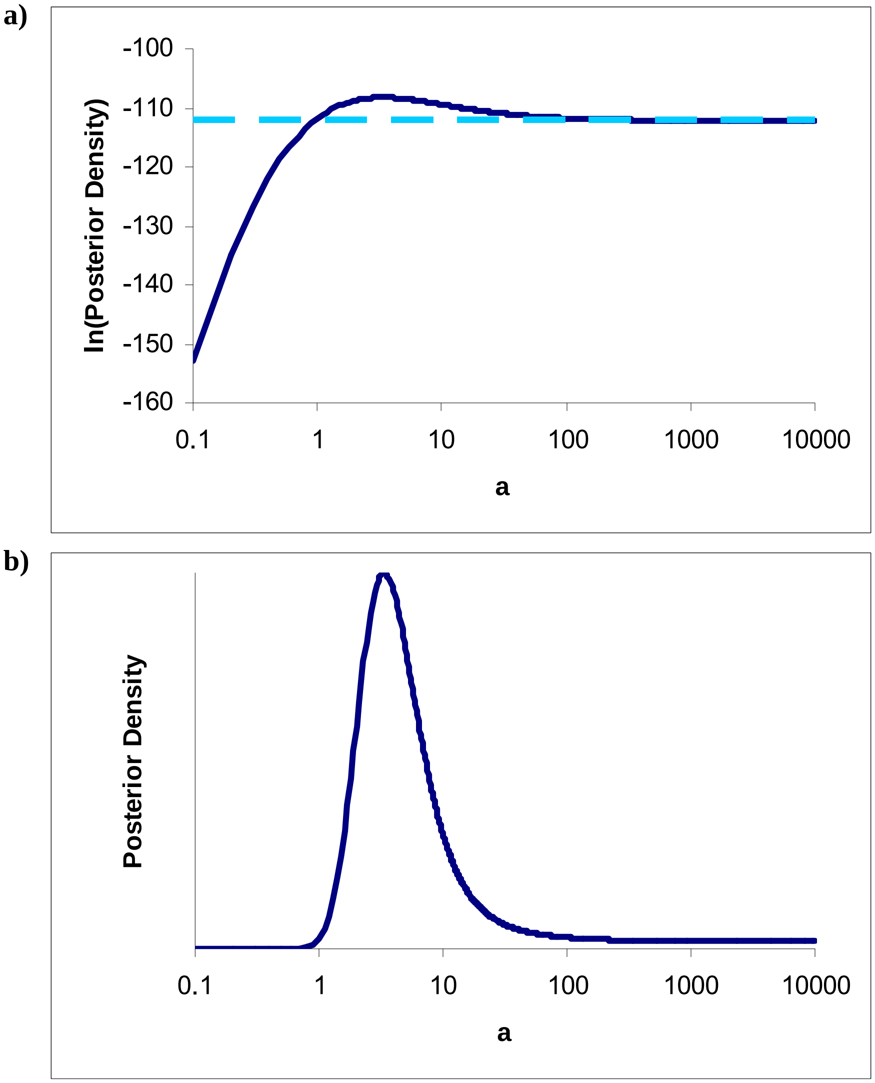


Figure 4.23: Shape of the posterior density function with improper uniform priors

These graphs show the shape of the posterior density function associated with the Table 4.6 data (using the beta-binomial recovery model and semi-infinite uniform priors) along the line $b = a \cdot (1/\mu_p - 1)$ with $\mu_p = 9.04\%$. Panel (a) shows the natural logarithm of the posterior density, which converges asymptotically upon the log-likelihood associated with the 'binomial-only' recovery model (in which the variability in analytical recovery among samples is zero). Panel (b) shows the shape of the posterior with the right tail that never converges upon zero.

$g(a,b) = exp(-(a + b) / 1000)$. When using exponential priors as relatively uninformative priors to ensure that the resulting posterior is proper, it is necessary to ensure that the supplied constant t ($t = 100$ and $t = 1000$ in these examples) is high enough that it will not have an undue impact upon the posterior distribution. A property of such exponential priors that may be useful in selecting an appropriate value of t is $P(a + b > 3t) \approx 0.05$. Setting t too high can result in a posterior that is locally flat and has been observed to cause poor mixing. Results of Gibbs sampling runs with these four alternative priors, 100,000 iterations, and no burn-in are shown in Figure 4.24. In general, these figures suggest that the Markov chains mix well when the beta distribution parameters are low and not when the parameters are high. This is caused to some extent by the logarithmic scale used for the parameter values: similar parameter value changes between iterations have been found at high and low values of the parameters, but these get smaller on a logarithmic scale as the parameter value increases. Many iterations are needed for the Markov chain to move from high parameter values to low parameter values, and so convergence will be slowest for posteriors that include many high parameter values. Figure 4.24b looks like it may be diverging at the end, but additional runs confirm that the Markov chain does return to typical parameter values.

The empirical cumulative distribution functions of the sequences that represent the posterior predictive distribution of analytical recovery and the posterior distributions of the mean and standard deviation of analytical recovery (results not shown) were compared to evaluate sensitivity of the results to the choice of prior. It was found that the curves representing the posterior predictive distribution of analytical recovery and the posterior distribution of mean analytical recovery were quite similar and that the exponential prior $g(a,b) = exp(-(a + b) / 1000)$ yielded the most different results (because it favours low beta distribution parameters values less so than the other considered priors). The curves representing the posterior distribution of the standard deviation of analytical recovery were quite different because each prior favours low beta distribution parameter values (which correspond to a high standard deviation of analytical recovery) to a different extent. These results demonstrate that a set of relatively uninformative priors can be quite uninformative about the mean or predicted analytical recovery values but can have an effect upon the posterior for the standard deviation of analytical recovery.

For the analyses considered in Sections 5.3.5 and 7.2, the improper prior $g(a,b) = (ab)^{-1}$ is used. There is some preliminary evidence that the resulting posterior is proper because Markov chains started with high initial parameter values consistently return to lower parameter values near the maximum likelihood estimates (which has not been found to be true for the prior $g(a,b) = (ab)^{-0.5}$). The dataset was analyzed using initial parameter estimates based upon the method of moments (as recommended in Section 4.4.3) and a burn-in of 30,000 iterations. Every 30th parameter pair of the following 900,000 iterations (which took close to an hour to generate) was recorded. Given satisfactory evidence of convergence from diverse initial parameter values as well as acceptable mixing, it would likely have been possible to obtain a reasonable approximation with a smaller burn-in and fewer iterations. The longer burn-in, longer run, and thinning (Gelman *et al.*, 2004) were used to yield better-converged results. The results of this Gibbs sampling run are plotted in Figure 4.25, which shows cumulative posterior density curves associated with the posterior distributions of the mean and standard deviation and the posterior predictive distribution of analytical recovery.

4.4.5 Testing Hypotheses about Analytical Recovery

It is often desirable to conduct some sort of statistical analysis upon the results of a recovery experiment to determine whether the analytical recovery of one set of data is higher or less variable than the analytical recovery of another set or to determine if the mean or standard deviation is better than some threshold value. Conventional parametric statistical tools such as t-tests, chi-squared tests,

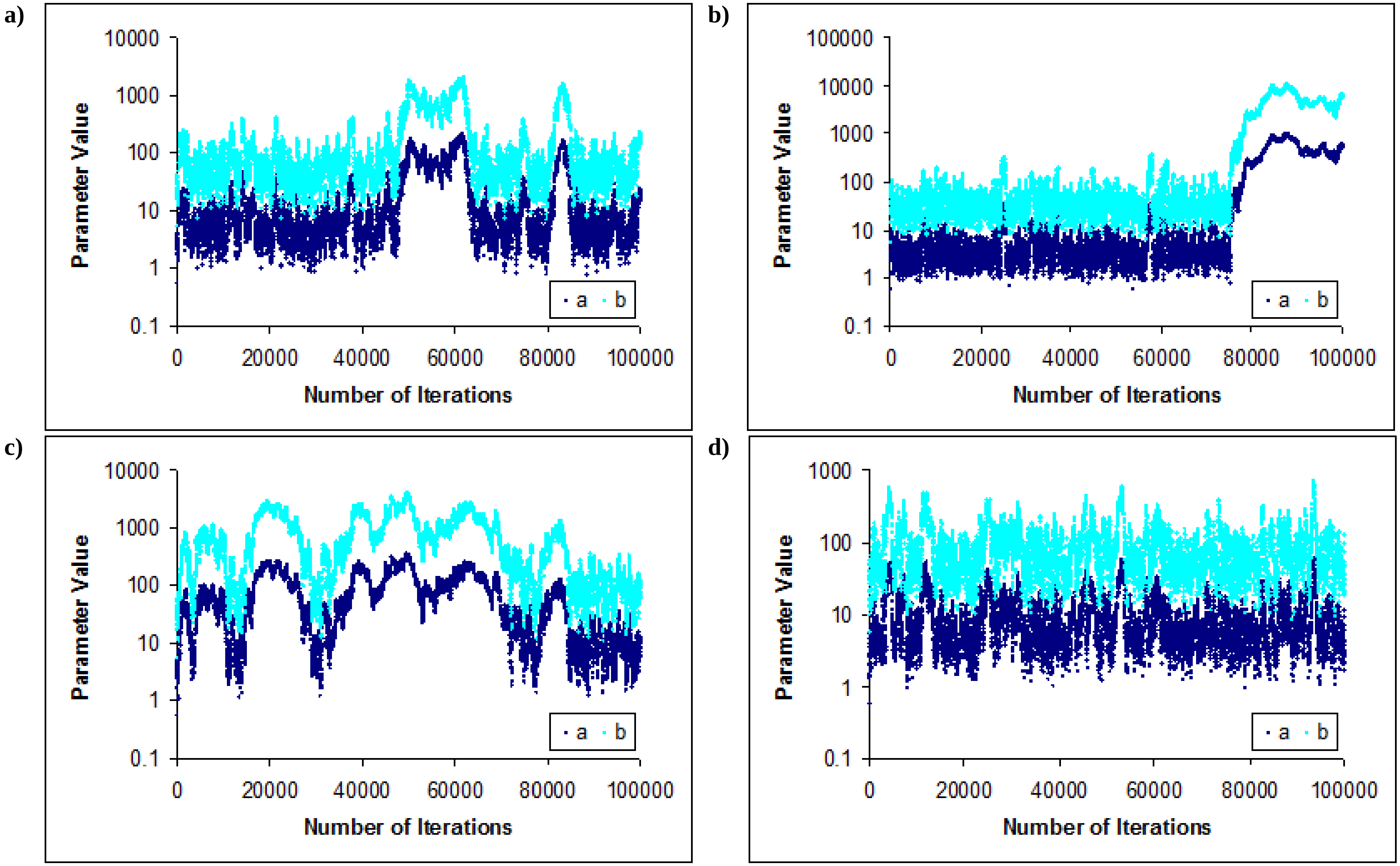


Figure 4.24: Results of Gibbs sampling using the Table 4.6 data, beta-binomial recovery model, and four alternative priors

These results were obtained without a burn-in and only every tenth pair of beta distribution parameters is plotted. The priors that were used are (a) $g(a,b) = (ab)^{-0.5}$, (b) $g(a,b) = (ab)^{-1}$, (c) $g(a,b) = exp(-(a + b) / 1000)$, and (d) $g(a,b) = exp(-(a + b) / 100)$.

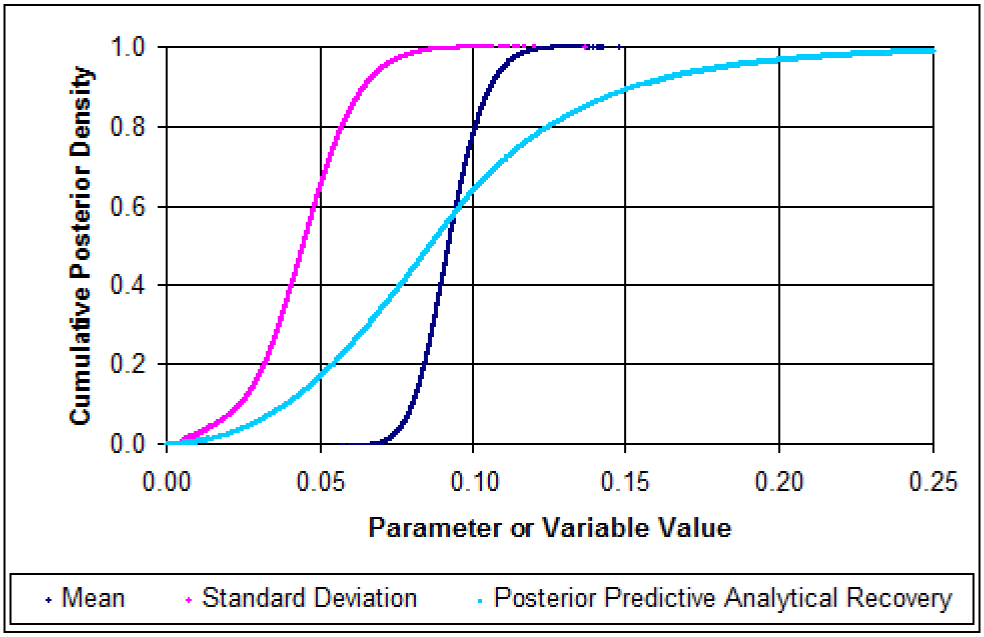


Figure 4.25: Posterior distributions for the Table 4.6 data using the prior $g(a,b)=(ab)^{-1}$

These cumulative posterior density curves were obtained using Gibbs sampling with a burn-in of 30,000 iterations, followed by a sequence of 900,000 additional iterations (which was thinned to every 30th iteration).

and F-tests (and their associated confidence intervals) are commonly used in such applications, but are subject to many limitations as discussed below. Non-parametric statistical tools, which are not addressed herein, may resolve some of the limitations associated with parametric statistics. Bayesian analyses (using specific probabilistic models) can also be used to compute credible intervals or posterior probabilities that represent the weight of evidence against a particular null hypothesis.

Limitations of conventional parametric statistical analyses

Common statistical tools include hypothesis tests, confidence intervals, and regression (which leads to ANOVA – analysis of variance). Each is based upon critical assumptions. For example, a t-test or a chi-squared test (Section 3.2.1) assumes that the data are independent, normally distributed, and homoscedastic (*i.e.* they have equal variance). ANOVA assumes that the error term in the model (which relates to the residuals) is normally distributed and homoscedastic. In some cases, a good approximation can be obtained despite small departures from these assumptions. In other cases, transformations can be used to yield data that more appropriately meet the assumptions of the statistical tools being used.

The fundamental problem with many of these statistical tools, even when the above assumptions are suitably met for a given dataset, is that they enable inference about the distribution of the data and not about the distribution of the parameter of interest. For example, it is not the mean and standard deviation of the recovery estimates that is of interest, but the mean and standard deviation of analytical recovery itself. It has been conclusively shown in Section 4.3 that even unbiased recovery estimates are imperfect measures of analytical recovery because of seeding and analytical error: the standard deviation of the recovery estimates can be substantially greater than the variability in analytical recovery itself. As a result, any statistical inference about the variance, standard deviation, or relative standard deviation of analytical recovery is biased high. The results of any such statistical analysis are only applicable to the recovery estimates (for a specific seeding method and quantity of seeded particles) and do not suitably address the variability in analytical recovery itself. Inferences about mean analytical recovery, on the other hand, are unbiased because the conventional recovery estimates are unbiased (as proven in Sections 8.1.1 and 8.1.2). The standard error of the mean, however, depends upon the standard deviation of the recovery estimates and is therefore inflated by excess variability due to seeding and analytical error.

Obtaining more precise recovery estimates by reducing the effects of seeding and analytical error (*e.g.* as described in Section 4.3) will on average result in smaller standard errors and will result in an increased power to detect small differences in mean analytical recovery. For example, the power of an upper-tailed z-test, with significance α, to successfully reject a null hypothesis (H_O: $\mu = \mu_O$) when the alternative hypothesis (H_A: $\mu > \mu_O$) is true can be calculated using Equation 4.34. In this equation, Φ^{-1}(.) is the standard normal cumulative distribution function, z_α is the standard normal test statistic with significance α, μ is the actual mean, μ_O is the null hypothesis mean, n is the number of data, and σ is the variance of the population of the data (after Snedecor and Cochran, 1989). The power is the complement of the probability of Type II error for the hypothesis test, which is decreased as σ is increased. If the population variance of the recovery estimates is decomposed into $\sigma^2 = \sigma_p^{\,2} + \sigma_e^{\,2}$ so that $\sigma_p^{\,2}$ represents the actual variance of analytical recovery and $\sigma_e^{\,2}$ represents the excess variability of the recovery estimates due to measurement errors (*i.e.* seeding and analytical error), then Equation 4.34 can be rewritten as Equation 4.35 (in which the power is decreased by increasing σ_e). Therefore, excess variability in recovery estimates due to measurement errors compromises the power of hypothesis tests to reject the null hypothesis when it is untrue. Accordingly, it behoves the analyst to minimize measurement errors where possible (*e.g.* as discussed in Section 4.3) and to explicitly state the quantity of seeded particles and probable error therein.

$$POWER = 1 - \Phi^{-1}\left(z_\alpha - (\mu - \mu_O) \cdot \frac{\sqrt{n}}{\sigma} \right) \tag{4.34}$$

$$POWER = 1 - \Phi^{-1}\left(z_\alpha - (\mu - \mu_O) \cdot \frac{\sqrt{n}}{\sigma_p} \cdot \sqrt{\frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2}} \right) \tag{4.35}$$

This discussion emphasizes the numerous limitations of conventional hypothesis tests that make it difficult to analyze recovery data correctly.

- Recovery estimates are unlikely to be normally distributed because skewed distributions are typical unless the mean is near 50% and the variance is low.
- Recovery estimates that are not fully replicated (*i.e.* data that do not share identical sample characteristics, seeding process, quantity of seeded particles, and sample

processing) will not have homogeneous errors because the seeding error and/or analytical error and/or non-constant analytical recovery will be inconsistent.

- Analytical recovery datasets are often too small (*e.g.* hypothesis tests in the literature are often based on sets of only 3 or 4 data) to enable selection of an appropriate and defensible transformation that enables valid conventional statistical analyses.
- The power of hypothesis tests to reject false null hypotheses is reduced by measurement errors in the recovery estimates.

If the foregoing limitations are ignored, then the ensuing results can be unreliable. It is, however, typically difficult to evaluate the reliability of published statistical analyses on analytical recovery because the assumptions are often not stated or validated, raw data are generally not provided, and the utilized statistical tools are often not explained in sufficient detail. Confidence intervals that exceed the parameter space (*e.g.* intervals that contain analytical recoveries below 0%) are strong evidence that statistical tools have been used without appropriate regard for their assumptions. For example, Rosen *et al.* (2009) considered hypothesis tests and confidence intervals upon the difference between two proportions (the proportion of unacceptable drinking water distribution system coliform results before and after some type of mitigation). They routinely computed confidence intervals on single proportions in which the lower bound was below 0%. These results were to be expected because the tool that was used is based upon a normality assumption that is only valid when the proportions are near 50% or the number of trials is very large (both of which were routinely violated). Such results are incorrect and compromise the reliability of all similarly conducted statistical analyses because the reader cannot determine which, if any, of the results correspond to acceptable assumptions. A simple diagnostic to check for problems with a normality assumption and the endpoints of the parameter space is to calculate the probability that the parameter of interest (*e.g.* proportion in the preceding example, analytical recovery here, and concentration in Section 5.5.1) lies outside its parameter space assuming normality. If analytical recovery is constrained to the interval $0 < p < 1$, for example, then the prediction interval represented by Equation 4.36 can be used.

$$1-\alpha = P\left(\frac{0-\overline{p}}{s_p\sqrt{1+1/n}} < T < \frac{1-\overline{p}}{s_p\sqrt{1+1/n}}\right) \tag{4.36}$$

In this equation, α is the probability that a future recovery estimate will fall outside the interval (0,1) given assumed normality of the recovery estimates and a dataset with n recovery estimates, sample mean $\overline{p}$, and sample standard deviation s_p. Assuming that analytical recovery cannot exceed 100%, the resulting value of α computed for the Table 4.4 recovery data is 0.00032, which is quite small. Therefore the validity of statistical analyses of this dataset that are based on normality is limited by the shape of the distribution of the data (*i.e.* whether or not it is actually normal) and the measurement errors in the recovery estimates, and not substantially by the endpoints of the parameter space. A normal scores plot (not shown) casts doubt upon the assumption of normality for the recovery estimates, but there are too few data to evaluate normality conclusively.

Bayesian analysis of analytical recovery

The two main advantages of a Bayesian analysis of analytical recovery are that a properly constructed probabilistic model accounts for the relationship between the recovery estimates and the actual analytical recovery that is of interest (*i.e.* it accounts for measurement errors) and that it facilitates incorporation of subjective information about the unknown parameters. A posterior distribution yields considerable information about the unknown parameters. For example, a 95% credible interval represents an interval within which the analyst is entitled to believe that the true parameter value lies with 95% probability given the model, available data, and selected priors. In contrast, a single 95% confidence interval makes no assertion about the possible values of the unknown parameter: however, 95% of such confidence intervals upon repetition of the experiment would contain the true parameter value. A posterior probability indicates what the analyst is entitled to believe about an unknown parameter while the significance in a hypothesis test only indicates the percentage of such hypothesis tests (if the experiment were repeated many times) that would falsely reject the null hypothesis when it is true. Improper posteriors can also be informative about the posterior density (*i.e.* the product of the prior and likelihood) of one parameter value versus another, but these do not have many of the convenient characteristics of proper posteriors. While classical frequentist analyses are often complicated by non-normality, non-constant variance, and failure to distinguish between the distribution of the parameter of interest and the distribution of the associated data, Bayesian statistics are complicated by the selection of a suitable prior and computational difficulties.

Once a proper posterior has been suitably evaluated (whether by integration or Markov Chain Monte Carlo methods such as Gibbs sampling), evaluation of hypotheses can be quite simple. For example, the posterior probability that the mean analytical recovery is less than 0.07 in Figure 4.25 is approximately 0.00695, which is compelling evidence that the mean is greater than 0.07. Similarly, the posterior probability that the standard deviation of analytical recovery is greater than 0.08 is approximately 0.01795, which is moderately compelling evidence that the standard deviation is less than 0.08. It can also be said that future values of analytical recovery are predicted to fall within the interval $0.00787 < p < 0.18502$ with a 95% probability. Given a suitable prior that results in a proper posterior and assuming that Gibbs sampling (Section 4.4.3) yields a suitably converged representation of the posterior distribution in a reasonable number of iterations and amount of computational time, the Bayesian approaches presented herein can be implemented to analyze the distribution of analytical recovery represented by a single dataset. For example, the posterior predictive distribution of analytical recovery could be compared to a validation criterion such as the probability that analytical recovery falls below a specific threshold.

Using Bayesian approaches to analyze differences between independent datasets is only slightly more complicated. The posterior for each dataset can be evaluated by Gibbs sampling and the posteriors can be compared by Monte Carlo (because the datasets are independent). This Monte Carlo process (based upon Gibbs sampling) is illustrated schematically in Figure 4.26. The posterior distributions of the mean and standard deviation of analytical recovery are generated by transformation from the beta (or gamma) distribution parameters generated in the Gibbs sampling sequence. The Bayesian analog of a t-test on the difference between two means would be the posterior distribution of the difference between the two means. This posterior distribution can be generated by repeatedly drawing a random value from the two posterior distributions of mean analytical recovery and computing their respective differences. Similarly, the Bayesian analog of an F-test on the ratio of two variances is the posterior distribution of this ratio. Posterior probabilities from these distributions can be used directly to describe the weight of evidence supporting a particular alternative hypothesis (*e.g.* that $\mu_1 - \mu_2 > \delta$). The biggest limitations to implementing these Bayesian methods are (1) that the probabilistic models addressing measurement errors in recovery estimates are based upon many assumptions that may be difficult to validate, (2) that the results can be highly subjective depending upon the choice of prior (and subjective statistical analyses are controversial), and (3) the methods require complicated computations that can be very time-consuming (particularly when Gibbs sampling converges slowly). Appropriate use of frequentist parametric (or non-parametric) statistical tools, if and only if their associated assumptions have been stated and suitably validated, may be a more viable alternative for analysis of recovery data.

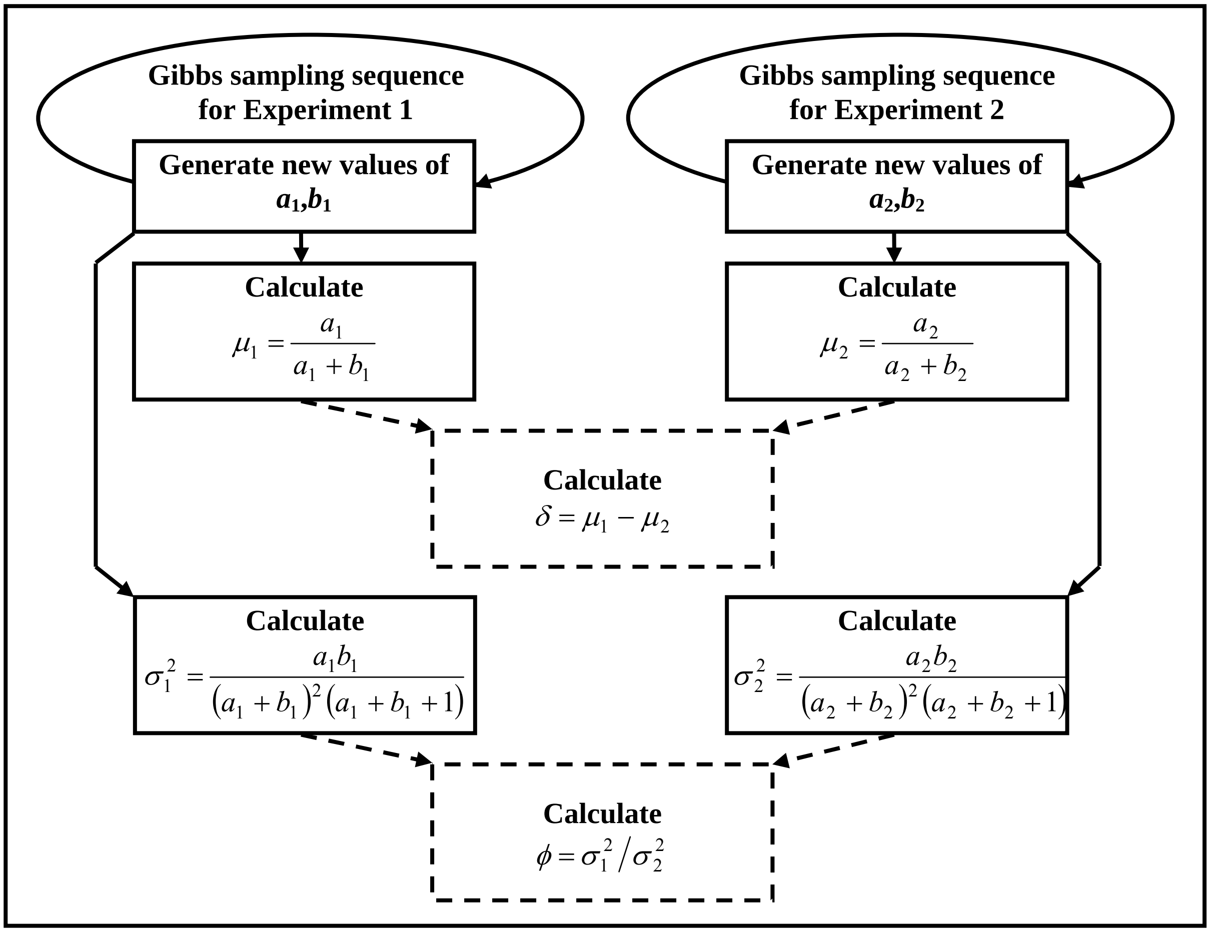

Gibbs sampling sequence for Experiment 1
Gibbs sampling sequence for Experiment 2
Generate new values of a_1,b_1
Generate new values of a_2,b_2
Calculate $\mu_1 = \frac{a_1}{a_1 + b_1}$
Calculate $\mu_2 = \frac{a_2}{a_2 + b_2}$
Calculate $\delta = \mu_1 - \mu_2$
Calculate $\sigma_1^2 = \frac{a_1 b_1}{(a_1 + b_1)^2 (a_1 + b_1 + 1)}$
Calculate $\sigma_2^2 = \frac{a_2 b_2}{(a_2 + b_2)^2 (a_2 + b_2 + 1)}$
Calculate $\phi = \sigma_1^2 / \sigma_2^2$

Figure 4.26: Evaluating the difference between two means or the ratio of two variances
The posterior distribution for the beta distribution parameters (a,b) in the beta-binomial recovery model is evaluated using the Gibbs sampling algorithm from Section 4.4.3. A sequence of mean and variance values is generated by transformation of the parameters a,b in each Gibbs sampling iteration. A sequence of values of the difference between the two means (δ) or the ratio of the two variances (ϕ) can be generated by Monte Carlo using randomized sequences of the means and variances.

4.5 PRACTICAL IMPLICATIONS

Analytical recovery is typically estimated as the count obtained from a seeded sample divided by the quantity of seeded analyte particles (*i.e.* microorganisms or discrete particles). This estimate is imprecise, particularly if the number of seeded particles is not precisely known. Quantitative analyses of recovery data must address these measurement errors to properly characterize the variability in analytical recovery. Accordingly, probabilistic models were developed herein to describe the count obtained as a random variable that depends upon the number of seeded particles (and possible variability therein) and upon the variability in analytical recovery among samples. These models were primarily used in two ways: to develop experimental design guidelines for recovery experiments and to develop quantitative analysis tools that address measurement errors in recovery estimates. A summary of specific contributions of this work that will enable better experimental quantification and statistical analysis of analytical recovery is provided below.

Strategies to quantify analytical recovery more effectively (Section 4.3)

- The count obtained from a seeded sample divided by the presumed number of seeded particles is an unbiased estimate of analytical recovery, but the estimate is imprecise due to random measurement errors. Therefore, the variability of a set of recovery estimates will be greater than the variability in analytical recovery itself

- It behoves analysts to design recovery experiments so that recovery estimates are precise. In general, this is accomplished by using precise numbers of seeded particles and increasing the number of seeded particles (if it is feasible and does not affect the representativeness of results).

- When conducting a recovery experiment with multiple replicate samples to quantify the mean and/or standard deviation of recovery, it behoves the analyst to choose a quantity of seeded particles and number of replicate samples that results in a precise estimate of the mean and an unbiased (and relatively precise) estimate of the standard deviation. A statistical tool was developed herein to facilitate analyses of experimental designs given preliminary estimates of the parameters of a beta distribution that describes variability in analytical recovery. This Monte Carlo simulation tool can be implemented to compare alternative experimental designs

- It was found, in general, that lower seed doses were acceptable when analytical recovery is more highly variable because recovery estimates will become sufficiently precise (*i.e.* to yield an accurate estimate of the standard deviation of analytical recovery) at a lower dose; conversely, when analytical recovery is less variable, higher seed doses are needed to obtain sufficiently precise recovery estimates.

- Although seeding samples with precisely known numbers of particles yields more precise recovery estimates than samples seeded with an aliquot of stock concentration, the benefit becomes negligible when mean analytical recovery is low.

- Seeding a sample with a precisely known number of analyte particles and then enumerating only a fraction of the sample partially (or in extreme cases wholly) negates the benefits of seeding samples with precisely known numbers of particles.
- The presented Monte Carlo method can also be used to evaluate the extent to which the standard deviation of the recovery estimates is likely to be an over-estimate of the standard deviation in analytical recovery.

Appropriate statistical analysis of replicate recovery data (Section 4.4)

- Statistical analyses that ignore measurement errors will over-estimate the variability in analytical recovery because the variability of recovery estimates is typically greater than the variability of recovery itself (due to measurement error).
- The mean and standard deviation of analytical recovery (or related model parameters) should be estimated using a probabilistic model that addresses measurement errors in the recovery estimates: maximum likelihood estimation can be used in this way.
- Bayes' theorem (and Gibbs sampling) were implemented to provide a quantitative description of uncertainty in parameters that describe variability in analytical recovery given a model that describes the measurement errors in the data.
- Conventional statistical confidence intervals and hypothesis tests are often inappropriate for analysis of recovery data because the data are often not normally distributed, may not have constant variance, and are often too few in number to enable transformation. Conventional analyses of the variance will be biased due to unaddressed measurement error in the recovery estimates. Analyses of the mean will not be biased, but the over-estimated variability will reduce the power of hypothesis tests to classify a difference from the null hypothesis as statistically significant. The proposed Bayesian approaches can be adapted to yield more robust analyses.

A very important conclusion from this work (that is certainly the easiest to address) is that reported recovery estimates that are not associated with information about the presumed number of seeded particles and the seeding method (or an appropriate measure of uncertainty in the recovery estimate) have little scientific value because the precision of the reported values cannot be assessed and it is not possible to subsequently conduct a proper statistical analysis of the data: it is imperative that information about seeding always be clearly provided along with any reported recovery estimates.

In summary, the analytical recovery of enumeration methods must be addressed to calibrate count data to physically relevant quantities of analyte particles, but there are also random measurement errors that affect the precision of analytical recovery estimates. Better analysis of enumeration methods can be accomplished by designing recovery experiments that yield more precise recovery estimates (so that optimally informative data are obtained using available resources) and by using statistical tools that properly address the random errors in the recovery estimates that are obtained.

Chapter 5

Accounting for Measurement Errors in Concentration and log-Reduction Data: Enabling Decisions Based upon Quantitative Measures of Uncertainty

“Particle and microbe enumeration data are inherently variable because (1) the concentration in the source may vary (representative sampling error), (2) replicate samples from a homogeneous source may contain different numbers of particles (random sampling error), (3) some of the particles in the sample may not be observed due to incomplete analytical recovery (analytical error), (4) analytical recovery may vary between equivalently processed samples (non-constant analytical recovery), or (5) repeated enumerations of a processed sample may yield different counts (counting error)” (Emelko *et al.*, 2010a). Accordingly, measurement errors must be considered in the analysis of enumeration-based concentration data because it is otherwise unclear whether the obtained concentration estimates are representative of (*i.e.* correctly calibrated to) actual concentrations and because it is unclear how much of the variability among potential non-replicate concentration estimates is due to measurement error and how much is due to actual temporal or spatial variation in concentration. The term ‘concentration estimate’ is used throughout this thesis rather than ‘concentration measurement’ because the raw measurements are counts that are subsequently used to estimate concentrations. The number of analyte particles observed in a sample of particular size is an estimate of concentration that is subject to several types of measurement errors.

The measurement errors that make concentration estimates imprecise, and also inaccurate if analytical recovery is not addressed, were introduced in Section 2.1. In this chapter, concentration heterogeneity (*i.e.* temporal or spatial variability in analyte particle concentration) is defined as representative sampling error because only individual samples and replicates are addressed (temporal concentration variability is addressed in Chapter 6). Individual samples have little value if they are not representative of the source that is being evaluated and repeated samples that are not representative of the same (locally) homogeneous source concentration cannot be appropriately regarded as replicates. Random sampling error describes the randomness of the number of analyte particles contained in a sample from a source with a locally homogeneous concentration because the particles are discrete, analytical error is the difference between the number of analyte particles observed in a sample and the number that are actually present, non-constant analytical recovery represents the variation in analytical recovery among samples, and counting error is the difference between the number of enumerable analyte particles in a processed sample and the count that is obtained by the analyst. Throughout this thesis, counting error is regarded as a component of analytical error (which is incorporated into information about analytical recovery) because analytical errors associated with sample processing and the errors in enumerating the processed sample are generally not considered or modelled separately. Proper analysis of concentration data must assess what concentration might have resulted in the counts, given knowledge about random measurement errors, rather than simply dividing counts by their respective sample volumes.

The most probable number (MPN) approach to estimate microbial concentrations from an array of presence-absence samples uses a simple probabilistic model to evaluate what the concentration might have been that resulted in the obtained data. It is generally understood that simple arithmetic operations cannot be used to infer the concentration from MPN data. The same is true for enumeration-based concentration estimates, yet counts per unit volume are often treated as precise concentration measurements without regard for uncertainty or bias. When analytical recovery is disregarded (unless it is on average 100%), decisions and models that use these enumeration-based concentration estimates can be biased because the concentration estimates are uncalibrated to actual concentrations. Even if the concentration estimates are unbiased (*e.g.* if they are divided by sample-specific recovery estimates or the mean analytical recovery that is representative of the enumeration method and type of sample analyzed), the estimates are uncertain, and repeated estimates will show some variability that is due to measurement error rather than actual variability in the particle or microorganism concentration. These issues will similarly affect log-reduction estimates associated with various treatment technologies (*e.g.* filtration, disinfection): log-reduction estimates can be biased if the analytical recovery associated with the two datasets is unequal and will be uncertain. It behoves analysts to consider measurement errors in enumeration-based concentration and log-reduction data so that experimental design can be enhanced to yield more informative data with the available resources and so that appropriate scientific value is obtained from data that are often costly and difficult to obtain. In particular, it is important to ensure that decisions intended to protect public health are made with appropriate regard for the bias and uncertainty in the concentration and log-reduction estimates upon which they are based.

The objectives of this chapter are as follows:

- to develop probabilistic models that describe random errors in concentration data (Section 5.1),
- to discuss and investigate assumptions surrounding the dispersion of microorganisms and discrete particles in suspensions (Section 5.2),
- to apply Bayes' theorem to evaluate the uncertainty in concentration estimates based upon individual and replicate enumeration data given a probabilistic model that addresses the random errors in the enumeration method (Section 5.3),
- to evaluate the uncertainty in individual log-reduction estimates by addressing the uncertainty in the concentration estimates upon which they are based (Section 5.4),

- to compare and contrast conventional confidence intervals and hypothesis tests with Bayesian approaches when analyzing enumeration-based concentration data (Section 5.5), and
- to investigate the factors that affect uncertainty in concentration and log-reduction estimates and to discuss strategies that may enable collection of more precise concentration and log-reduction estimates (Section 5.6).

The chapter is concluded with a discussion of practical implications (Section 5.7). Several case studies and illustrative examples are provided throughout this chapter to exemplify various concepts and to illustrate use of the proposed quantitative approaches.

5.1 MODELLING RANDOM ERRORS IN CONCENTRATION DATA

The conventional concentration estimate associated with enumeration-based methods is the number of analyte particles observed in a sample divided by the sample volume. Unavoidable measurement errors, however, cause these concentration estimates to be imprecise and also inaccurate if the analytical recovery is not 100%. Interpretation of concentration data without consideration of measurement errors can be inappropriate because the concentration estimates must be calibrated to actual concentrations using knowledge about the analytical recovery of the enumeration method and because the concentration estimates may be quite uncertain. Consequently, it is necessary to develop models that describe the effects of various random errors upon enumeration data.

Herein, two probabilistic models (the ‘beta-Poisson enumeration model’ and the ‘negative binomial enumeration model’) are developed to describe the effects of random sampling error, analytical error, and non-constant analytical recovery upon replicate enumeration data. Replicates are defined herein as samples that are representative of the same locally homogeneous source concentration. They do not necessarily have equal sample volumes nor do they need to have been obtained using identical sample processing and enumeration methodology. The two models are based upon suitable mathematical assumptions and stochastic processes and are therefore generally applicable without the need for case-specific fitting (provided that the stated assumptions are reasonably met). The main difference between the two is associated with analytical recovery. Recovery is regarded as a probability (that cannot exceed 100%) in the beta-Poisson enumeration model (Section 5.1.1) while it is regarded as a rate (that can exceed 100% due to random counting errors) in the negative binomial enumeration model (Section 5.1.2). These models are equivalent to the beta-Poisson and negative binomial recovery models (Sections 4.2.3 and 4.2.4) except that seeding error is replaced with random sampling error and concentration is the unknown parameter of interest rather than the parameters of the non-constant analytical recovery distribution. Accordingly, both models require information about the variability in analytical recovery among samples, which can be quantified and modelled as described in Chapter 4. Discussion of minor modifications to these models to address partial sample analysis or include sample-specific recovery information is discussed in Section 5.1.3. The presented models are contrasted with other probabilistic models for concentration data in Section 5.1.4.

5.1.1 The Beta-Poisson Enumeration Model

The beta-Poisson enumeration model is based upon a model presented by Nahrstedt and Gimbel (1996) that was also further developed by Emelko (2001). A more complete discussion of the model and its assumptions was provided by Emelko *et al.* (2010a). The model assumes that random sampling error is Poisson-distributed, analytical error is binomially distributed, and non-constant analytical recovery is beta-distributed. It also assumes that analytical recovery is independent of the concentration and sample volume (or at least that the available recovery information is adequately representative of relevant concentrations and sample volumes). If it is assumed that particles are randomly distributed throughout a locally homogeneous source (*i.e.* a source that has a constant concentration throughout the region and segment of time from which samples assumed to be replicates were collected) and that samples are independent, then the number of particles in a sample of specified volume is Poisson-distributed. This has been mathematically proven, and also empirically demonstrated for yeast cells upon a hemocytometer slide, by ‘Student’ (1907). If the particles are not randomly distributed throughout the source (*e.g.* they are clumped, settled, or attached to other particulate), then the Poisson assumption is not appropriate. Non-random dispersion is discussed in greater detail in Section 5.2. If a source has a finite volume, then a binomial model for sampling error may be more appropriate and repeated samples may not be independent as discussed in Section 4.2.3.

The binomial model for analytical error is based upon the assumption that each particle present in a particular sample has an equal probability of being observed (*i.e.* analytical recovery) and that the observation of any particle is independent of the observation or loss of the other particles (Section 4.2.2; Schmidt *et al.*, 2010a; Emelko *et al.*, 2010a). This binomial model precludes the possibility that analytical recovery can exceed 100% due to counting errors. The beta model that is used to describe non-constant analytical recovery is selected for mathematical convenience rather than being the result of specific assumptions and processes. It has been used extensively to model variability in analytical recovery among samples (*e.g.* Nahrstedt and Gimbel, 1996; Teunis *et al.*, 1997; Teunis *et al.*, 1999; Teunis and Havelaar, 1999; Crainiceanu *et al.*, 2003; Pouillot *et al.*, 2004; USEPA, 2005; Signor and Ashbolt, 2006; Petterson *et al.*, 2007; Jaidi *et al.*, 2009; Cummins *et al.*, 2010; Schmidt *et al.*, 2010a) “because it is confined between zero and one, is practically unimodal (*i.e.* beta distributions that are not unimodal are not of practical interest), and is the conjugate of the binomial distribution” (Emelko *et al.*, 2010a).

The resulting probabilistic model is represented by Equation 5.1, in which x is the number of observed particles in a sample of volume V taken from a source with homogeneous particle concentration c, n is the unknown number of particles that were actually present in the sample, p is the analytical recovery of the sample, and a,b are parameters of the beta distribution describing non-constant analytical recovery. The beta distribution parameters can be obtained from the results of a recovery experiment (Section 4.4.1, Schmidt *et al.*, 2010a). This model is called the beta-Poisson model because the Poisson and binomial components of the model can be combined into a single thinned Poisson model by summing out the nuisance parameter n (Equation 5.2). The marginal distribution of the number of observed particles (x) can be obtained by integrating the probability of recovery (p) out of Equation 5.2, but the integral cannot be solved explicitly (Equation 5.3).

$$f(x,n,p\,|\,c,V,a,b)=\left[\frac{e^{-cV}(cV)^{n}}{n!}\right]\left[\frac{n!}{x!(n-x)!}p^{x}(1-p)^{n-x}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}\right] \tag{5.1}$$

$$f(x,p\,|\,c,V,a,b)=\left[\frac{e^{-cVp}(cVp)^{x}}{x!}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}\right] \tag{5.2}$$

$$f(x\,|\,c,V,a,b)=\frac{(cV)^{x}}{x!}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\int_{0}^{1}e^{-cVp}p^{x+a-1}(1-p)^{b-1}dp \tag{5.3}$$

5.1.2 The Negative Binomial Enumeration Model

The negative binomial enumeration model is an alternative to the beta-Poisson model in which analytical recovery is regarded as a rate (rather than a probability) that can exceed 100% due to counting errors. In this model (Equation 5.4), random sampling error and analytical error are jointly modelled by a Poisson distribution (similar to Equation 5.2) and non-constant analytical recovery is modelled by a gamma distribution. Development of this model was motivated by Margolin *et al.* (1981) who modelled replicate plate-counts of *Salmonella* using a Poisson distribution with a mean that was itself gamma-distributed due to measurement errors. The model used herein, however, is more flexible because it separates the mean of the Poisson distribution into three parts: the particle concentration in the source (*c*), the sample volume (*V*), and the analytical recovery of the enumeration method (*p*). A model in which a Poisson-distributed random variable has a mean that is itself gamma-distributed will result in a negative binomial distribution (Fisher, 1941) as discussed in Section 3.1.2. The negative binomial marginal distribution for the number of observed particles (Equation 5.5) is obtained by integrating analytical recovery out of Equation 5.4 (Emelko *et al.*, 2010b).

$$f(x,p\,|\,c,V,\alpha,\beta)=\left[\frac{e^{-cVp}(cVp)^{x}}{x!}\right]\left[\frac{1}{\beta^{\alpha}\Gamma(\alpha)}p^{\alpha-1}e^{-p/\beta}\right] \tag{5.4}$$

$$f(x\,|\,c,V,\alpha,\beta)=\frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)}\left(\frac{cV\beta}{cV\beta+1}\right)^{x}\left(\frac{1}{cV\beta+1}\right)^{\alpha} \tag{5.5}$$

5.1.3 Models with Partial Sample Analysis and Sample-specific Recovery Information

The probabilistic models presented in Sections 5.1.1 and 5.1.2 can be expanded to include partial sample analysis or case-specific recovery information. Partial sample analysis, enumeration of only a portion of a sample, is discussed in Section 4.2.5. If the sample is well-mixed when it is subdivided, then the analytical error in enumerating the sample is related to the product of the fraction of sample analyzed (θ) and analytical recovery (*p*). Modification of the beta-Poisson and negative binomial enumeration models to include partial sample analysis yields Equations 5.6 and 5.7. It is clear that these models are equivalent to the beta-Poisson and negative binomial enumeration models presented above except that the volume of the sample is replaced with θV: the 'effective volume' (Parkhurst and

Stern, 1998). If a fraction θ of a well-mixed sample of volume V is enumerated, then the measurement error is equivalent to enumerating a sample of volume θV. Young and Komisar (1999) address the use of a negative binomial model for sub-sampling error when particles are not randomly distributed throughout the sample (*i.e.* when they are clumped). Such a model is impractical, however, as discussed in Section 4.2.8. It is imperative that samples be well-mixed during sub-sampling and it may be preferable to collect smaller samples so that sub-sampling is not necessary.

$$f(x,p\,|\,c,V,\theta,a,b)=\left[\frac{e^{-cV\theta p}(cV\theta p)^{x}}{x!}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}\right] \tag{5.6}$$

$$f(x,p\,|\,c,V,\theta,\alpha,\beta)=\left[\frac{e^{-cV\theta p}(cV\theta p)^{x}}{x!}\right]\left[\frac{1}{\beta^{\alpha}\Gamma(\alpha)}p^{\alpha-1}e^{-p/\beta}\right] \tag{5.7}$$

The models presented in this chapter assume that all replicates are processed using similar methodology so that the parameters for the non-constant analytical recovery distribution (a,b or α,β) are the same for all samples. This is not necessarily true because enumeration data that represent the same source concentration but that are obtained using different methodology should be analyzed together to make inferences about the source concentration. Any of the models presented in this chapter can be modified to include sample-specific recovery distribution parameters by adding the subscript i to represent the parameters for the i^{th} sample (Emelko *et al.*, 2010b). Such an expanded model simply allows the analyst to supply more detailed recovery information if it is available. This type of notation is used in Section 6.1.1.

In the enumeration of *Cryptosporidium* oocysts and *Giardia* cysts in natural water matrices, sample-specific recovery estimates may by obtained by seeding samples with precisely known numbers of uniquely labelled (oo)cysts that are enumerated separately from the indigenous (oo)cysts (*e.g.* Section 4.2.7; Warnecke *et al.*, 2003; Petterson *et al.*, 2007). If analytical recovery is regarded as a probability, then the number of seeded (oo)cysts that are observed (x^*) is binomially distributed as a function of the number of seeded (oo)cysts (n^*) and the 'internal seed enumeration model' becomes Equation 5.8 (in which p must be specified). If analytical recovery varies randomly among samples, then a non-constant analytical recovery distribution can be added to Equation 5.8 so that p is a random variable. Discussion of how to incorporate internal seed recovery data into Bayesian analysis of replicate concentration data is discussed in Section 5.3.4.

$$\begin{gathered}f(x,x^*,n\,|\,c,V,n^*,p)=\\ \left[\frac{e^{-cV}(cV)^{n}}{n!}\right]\left[\frac{n!}{x!(n-x)!}p^{x}(1-p)^{n-x}\right]\left[\frac{n^*!}{x^*!(n^*-x^*)!}p^{x^*}(1-p)^{n^*-x^*}\right]\end{gathered} \tag{5.8}$$

5.1.4 Other Probabilistic Enumeration Models

Use of the Poisson and negative binomial distributions to describe variability in enumeration data is relatively common. Indeed, use of the Poisson distribution is inherent to the analysis of data obtained using most microbiological enumeration methods (*e.g.* Eisenhart and Wilson, 1943). The most probable number (MPN) method of estimating microbial concentrations is also based on a probabilistic model in which random sampling error (which relates to the probability that a presence-absence sample will contain at least one target microorganism) is Poisson-distributed and the number of positive results in a number of equivalent presence-absence samples is binomially distributed (*e.g.* Haas and Heller, 1988). Negative binomial distributions have been widely used to describe variability in enumeration data when the data are over-dispersed with respect to the Poisson distribution (as discussed further in Section 5.2). Often, probabilistic models are chosen on the basis of case-specific fit without particular consideration to the random errors that may cause enumeration data to follow a particular type of distribution. This section specifically addresses probabilistic models that describe random errors in the enumeration process. Several models that are specific to the enumeration of (oo)cysts are summarized and contrasted with the models presented in this chapter, although the same concepts apply to enumeration of any type of microorganisms or discrete particles (unless analytical recovery is consistently 100%).

The beta-Poisson enumeration model presented in Section 5.1.1 is based upon a model proposed by Nahrstedt and Gimbel (1996). The assumptions associated with the distributions in their model are more fully explained in Emelko *et al.* (2010a) and throughout this thesis. Nahrstedt and Gimbel (1996) used the method of moments to estimate parameters for the beta distribution describing non-constant analytical recovery. It was demonstrated in Schmidt *et al.* (2010a) and in Section 4.4.1 that the method of moments yields biased parameter estimates because it does not account for seeding and analytical error in the recovery data; maximum likelihood estimation using a suitable probabilistic model for the recovery data will yield more appropriate point estimates of the parameters. Nahrstedt and Gimbel (1996) presented an approach to calculate confidence intervals representing a range of (oo)cyst concentrations that might have led to a particular count. Their method, however, has not been widely adopted because it is mathematically cumbersome and does not conveniently provide a full description of uncertainty in concentration. Furthermore, their model applied only to single enumerations (*i.e.* not to sets of counts that represent the same source concentration). Use of Bayes' theorem to obtain a posterior distribution (for single or replicate enumeration data) that more completely describes uncertainty in a concentration estimate than a confidence interval is presented in Emelko (2001), Emelko *et al.* (2010a), and Section 5.3.

Parkhurst and Stern (1998) presented a model that does not account for non-constant analytical recovery but that explicitly accounts for partial sample analysis. The models presented in Section 5.1.3 account for partial sample analysis (assuming that the sample is well-mixed when subdivided) in a similar way, but also address non-constant analytical recovery. These models are shown to be equivalent to substituting the effective volume into the models that do not account for partial sample analysis. Nahrstedt and Gimbel (1996) had stated that "if the sample is not counted out completely, the counted volume equivalent should be used for *V*." Accordingly, the model presented by Parkhurst and Stern (1998) is only a simplification of Nahrstedt and Gimbel (1996).

Young and Komisar (1999) used the model presented by Nahrstedt and Gimbel (1996) and considered partial analysis of the resuspended pellet following the centrifugation step. They considered Poisson and negative binomial models for random errors in sub-sampling. As discussed in greater detail in Section 4.2.8, their model is not mathematically appropriate because the sub-sampling error in a well-mixed sample is binomially distributed (which is approximately Poisson-distributed if the number of (oo)cysts is large and the enumerated fraction is small). Furthermore, their negative binomial model is impractical because it is a fitted model for which sample-specific parameter estimates cannot be obtained.

Teunis and Havelaar (1999) used the beta-Poisson model with an added gamma distribution describing temporal variability in source concentration. Similarly, Petterson *et al.* (2007) addressed temporal concentration variability using a gamma distribution and addressed random sampling and analytical error using a Poisson distribution. Three alternative models were used to represent varying amounts of information about analytical recovery: one in which analytical recovery is ignored (*i.e.* it is implicitly assumed to be consistently 100%), one in which analytical recovery is assumed to vary among samples according to a beta distribution, and one in which sample-specific recovery estimates obtained by internal seeding are available for each sample. Temporal concentration variability is addressed in Chapter 6, and further discussion of how these models were used to evaluate temporal concentration variability is included in Section 6.1.2.

Crainiceanu *et al.* (2003) used a Bayesian generalized linear mixed model to analyze nationwide (oo)cyst concentration data collected in accordance with the United States' Information Collection Rule. Their model accounted for random sampling and analytical error by using a Poisson distribution and accounted for non-constant analytical recovery with a secondary Bayesian generalized linear mixed model that was discussed in Sections 4.1.5 and 4.2.8. The model implicitly accounted for partial sample analysis by using the volume of water analyzed (*i.e.* the effective volume). A regression-based component was also used to investigate correlation between (oo)cyst concentrations and possible covariates (*e.g.* turbidity, temperature, pH, total coliform counts, season, population served by a water supply, reservoir residence time) and the residuals were partitioned into time-site effects, site effects, and regional effects. Such a complicated model can yield useful information about correlation (or lack thereof) between (oo)cyst counts and possible covariates, but data for possible covariates are often not available (or relevant when analyzing a set of replicate concentration data as is the focus in this chapter) and such regression models are unlikely to be predictive for other datasets or scenarios (because correlation is case-specific and easily obscured by other sources of variability).

5.2 DISPERSION OF MICROORGANISMS AND DISCRETE PARTICLES IN WATER

The models presented in this thesis assume that the microorganisms or discrete particles of interest are randomly distributed throughout the source with a locally homogeneous concentration (*i.e.* the concentration may vary over space or time but is consistent within the time and space from which samples regarded as replicates are collected). If repeated samples are independent (which is a reasonable assumption provided that the total sample volume is small in comparison to the volume of the source as discussed in Section 4.2.3), then the number of particles contained in a sample (*n*) will be Poisson-distributed with a mean equal to the product of concentration (*c*) and sample volume (*V*). Figure 5.1 illustrates this random sampling error.

The variance of the Poisson distribution is necessarily equal to the mean. The Poisson distribution, however, has often been criticized because the variance of microorganism counts is commonly found to be greater than the mean (*i.e.* the counts are over-dispersed relative to the Poisson distribution). This section includes a summary of literature that addresses microorganism dispersion in water and a discussion of the possible causes of over-dispersion (Section 5.2.1), the results of an experiment used to investigate dispersion (Section 5.2.2), and a discussion of the validity of the Poisson assumption in probabilistic models for random sampling error (Section 5.2.3).

5.2.1 Over-dispersed Count Data

If repeated samples from a source yield counts with a variance that is greater than the mean, then the data are over-dispersed relative to the Poisson distribution. Conversely, counts with a variance lower than the mean are under-dispersed. Over-dispersion can be indicative of a clustered particle distribution in the source; samples that contain large clusters will yield higher counts and samples that

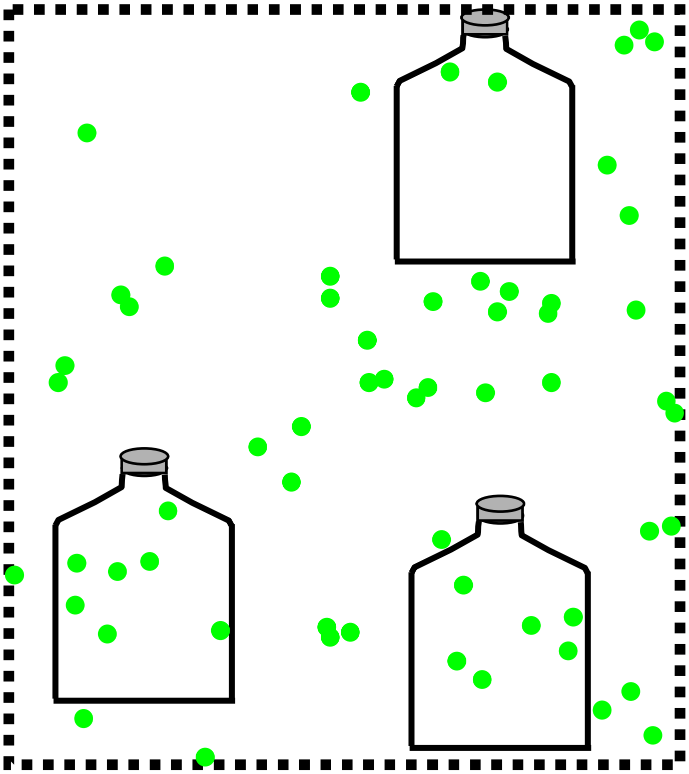


Figure 5.1: Random sampling error from a well-mixed source
(after Nahrstedt and Gimbel, 1996; Emelko, 2001)
When particles are randomly distributed throughout a source with constant concentration (*c*), repeated samples will contain different numbers of particles even if the sample volume (*V*) is constant.

do not contain large clusters will contain lower counts so that the overall variance will be greater than it would be if the particles were randomly distributed. Under-dispersion can be indicative of a systematic or somewhat evenly spaced particle distribution (*e.g.* if particles are mutually repulsive such as large trees in a forest). Even if the particles are randomly dispersed, the sample variance can be randomly less than or greater than the sample mean. Accordingly, a hypothesis test (Section 3.2.1) based on Fisher's index of dispersion (Fisher *et al.*, 1922) is commonly used to determine if the dispersion of a set of counts is significantly different from that of a Poisson distribution (Eisenhart and Wilson, 1943; El-Shaarawi *et al.*, 1981; Margolin *et al.*, 1981; Christian and Pipes, 1983; Haas and Heller, 1986; Gale *et al.*, 1997).

The negative binomial distribution is often proposed to fit over-dispersed counts because its variance is greater than its mean and it converges upon a Poisson distribution as its variance approaches its mean (Section 3.1.2). Various goodness-of-fit tests (*e.g.* chi-squared, Kolmogorov-Smirnov, likelihood ratio) have been used to evaluate the fit of Poisson, negative binomial, and other distributions (*e.g.* Poisson-plus-added-zeros, log-normal) to sets of counts (Pipes *et al.*, 1977; El-Shaarawi *et al.*, 1981; Christian and Pipes, 1983, Medema *et al.*, 1995; Haas and Rose, 1996; Gale *et al.*, 1997; Teunis *et al.*, 1997). In many cases, the negative binomial distribution will yield a superior fit to Poisson-distributed data simply because it has two fitted parameters rather than one. Accordingly, the negative binomial distribution should not be used on the basis of better fit than the Poisson distribution unless it has first been demonstrated (*e.g.* using the index of dispersion test) that the counts are significantly over-dispersed. The Poisson-plus-added-zeros distribution (in which the count is Poisson-distributed some fraction of the time and zero the rest of the time) is advocated when data-sets have a disproportionate abundance of non-detects. This is also a two-parameter distribution and may yield a better fit than the Poisson distribution for data that are actually Poisson-distributed (if there are many zeros). The log-normal distribution is an unacceptable choice for count data for two reasons (Gale *et al.*, 1997). First, discrete counts cannot be appropriately modelled by a continuous distribution (although the approximation may be good if the counts are high). The more important reason is that the log-normal distribution is incompatible with non-detects and the analyst must manipulate zeros into positive values (which introduces bias) in order to fit the distribution.

Haas and Heller (1988) explain that the argument against use of the Poisson distribution for random sampling error "has been based on studies of samples taken over time and has ignored the enumeration methodology itself as a source of variability".

> "It is important to separate the question of frequency distribution between replicates of the same sample from the distribution characterizing the temporal variability. If the former distribution is not Poisson, it may indicate that a deficiency in methodology exists preventing truly Poisson enumeration from occurring. The latter frequency distribution, however, may depend on the degree to which the enumeration technique is capable of producing results with Poisson variability, as well as on the intrinsic temporal variability of the microorganism density." – (Haas and Heller, 1986)

Margolin *et al.* (1981) considered a model in which the mean of the Poisson distribution varied among replicate plate counts according to a gamma distribution (*i.e.* each count is Poisson-distributed with a different gamma-distributed mean) and that this variation was due to inconsistencies in the methodology. The result was a negative binomial distribution (Section 3.1.2). Emelko *et al.* (2010a) stated that over-dispersion "may be due to (1) representative sampling error (*i.e.* the samples are not replicates), (2) non-random distribution of the particles (*e.g.* the particles are clumped or attached to other particulate), or (3) methodological errors that contribute variability to the data".

Both the condensed beta-Poisson model (Equation 5.2) and the negative binomial model (Equation 5.4) presented herein and in Emelko *et al.* (2010a) model counts as Poisson-distributed with mean cVp. Variation in any of these parameters (*i.e.* temporal or spatial variation in particle concentration, variation in the enumerated volume among samples, or non-constant analytical recovery of the enumeration method) will result in over-dispersed counts. In the case of replicate concentration data, there is no variability in concentration, variability in sample volumes is addressed, and random inconsistencies in the enumeration method are addressed by non-constant analytical recovery. Non-random particle distribution in an otherwise homogeneous source (*e.g.* clumping) is the only cause of non-Poisson counts that is not addressed in these models. Non-random particle distributions in the source, however, cannot be asserted solely on the basis of statistically significant under- or over-dispersion unless it can be proven that the over-dispersed counts did not arise from the other possible sources of variability. "Microscopic inspection of the environmental distribution of the particles in an unmixed sample may be beneficial [to investigate clumping], but the particle concentration is often impractically low and sample preparation may disrupt the environmental distribution of the particles" (Emelko *et al.*, 2010a). Section 5.2.2 provides an example of microscopic inspection of particle dispersion. The validity of the Poisson assumption for random sampling error in the models used in this chapter (and in Chapter 6) is discussed in Section 5.2.3.

5.2.2 Case Study: Point Pattern Analysis of Microspheres on a Hemocytometer

In spatial statistics, point pattern analysis is used to determine whether objects are under-dispersed, randomly distributed, or clustered. Spatial analysis of microorganisms or discrete particles in water, however, is complicated because it is not possible to obtain the natural three-dimensional coordinates of the particles and because the particles are unlikely to be stationary due to settling and currents. Methods used to enumerate or visually inspect samples will also generally disrupt the natural distribution of particles in the sample.

An experiment was conducted to investigate the possible presence of clumps in a stock suspension of 4.5 μm polystyrene microspheres following vortexing. A Petroff-Hausser counting chamber (Hausser Scientific, Horsham, PA) with a 0.02 mm depth was filled with a sample of the suspension. Based upon the resulting count, the stock concentration was estimated to be approximately 7.0×10^6 microspheres/mL. A total of nine digital images (each representing a 1 mm × 1 mm section of the 3 mm × 3 mm improved Neubauer grid) were captured using a 10x objective lens. These images (*e.g.* Figure 5.2) were then used to count the microspheres within each quadrat (*i.e.* rectangular section) of the grid. Point pattern data were obtained from the digital images using the centre coordinates of each microsphere (to the nearest half pixel). The coordinates were then transformed to represent the position of each microspheres relative to the grid. The counts from equal-sized quadrats were analyzed using Fisher's index of dispersion (Section 3.2.1) to determine if the counts were significantly under- or over-dispersed relative to the Poisson distribution. Several point pattern analysis tools in R (a free language and environment for statistical computing that is available at www.r-project.org) were used to investigate the distribution of microspheres on the grid.

The count data are displayed in Figure D.1 (in Appendix D.2), which also shows the approximate layout of the grid. Table 5.1 shows the results of two-tailed hypothesis tests using Fisher's index of dispersion for each of the nine 1mm × 1mm squares in the grid and for the composite counts across the grid (using two quadrat sizes). The sample variances are sometimes slightly larger and sometimes slightly smaller than the corresponding sample means (which is not strong evidence of clumping). The lowest computed P-value is approximately 0.174, indicating that none of the sets of counts are significantly under- or over-dispersed relative to the Poisson distribution.

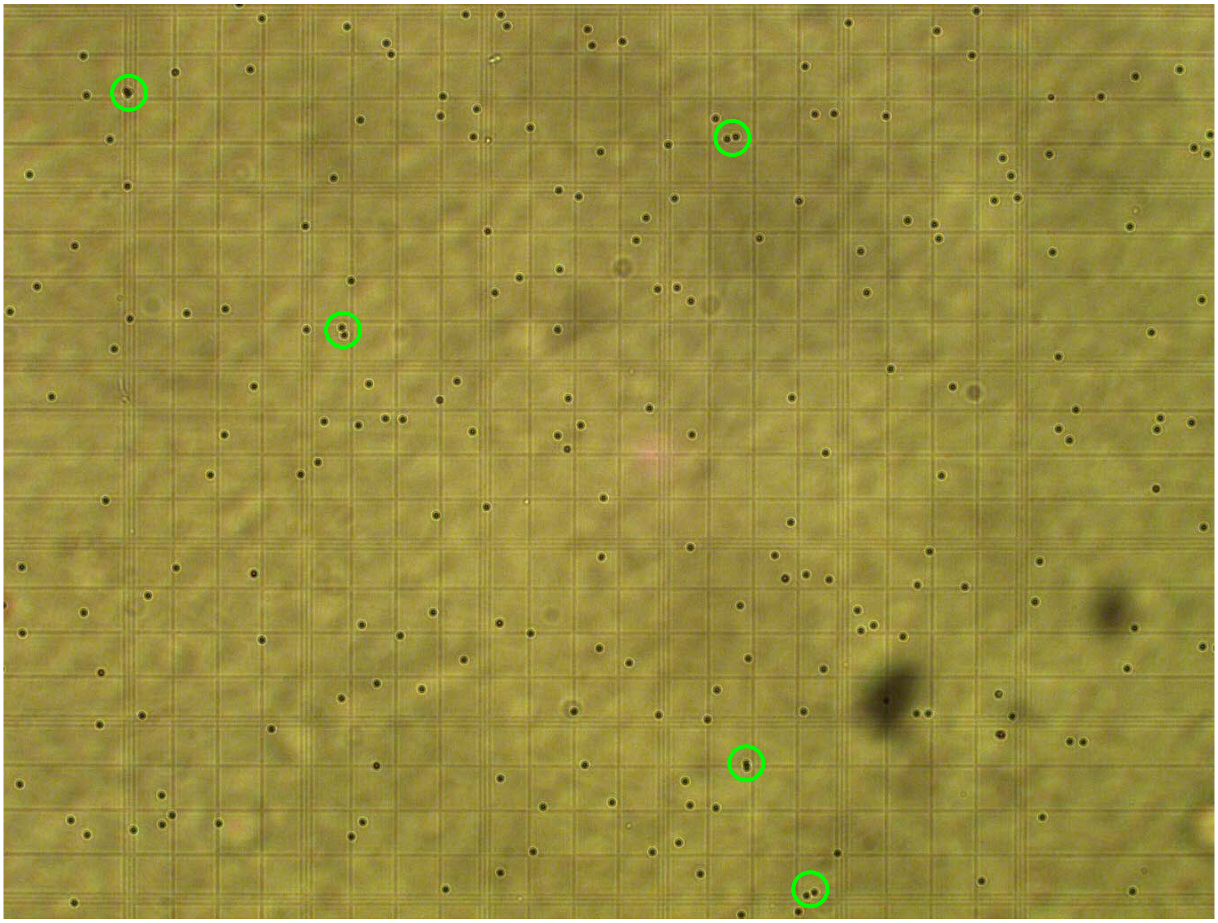


Figure 5.2: Polystyrene microspheres in a Petroff-Hausser counting chamber
Possible clumps of two or three microspheres are highlighted with circles.

Use of the Poisson distribution to model random sampling error in the enumeration of particles or microorganisms in water (or on a hemocytometer) assumes complete spatial randomness (*i.e.* that the particles are randomly distributed with homogeneous concentration). Obtaining counts that are not significantly under- or over-dispersed with respect to the Poisson distribution, however, does not prove complete spatial randomness because it is possible for counts to be Poisson-distributed at one scale (*i.e.* sample volume) and not at another. The index of dispersion tests summarized in Table 5.1, for example, fail to indicate significant differences from complete spatial randomness for their respective quadrat sizes, but do not preclude the possibility of clumping if smaller or larger quadrats had been used. Ripley's K-function (Venables and Ripley, 2002) can be used to address this issue. This function can be written as Equation 5.9, in which λ is the density (*e.g.* microspheres/μm^2) and $E[X(h)]$ is the expected value of the number of events (*e.g.* microspheres) within radius h of any given event (excluding the event at the centre). The expected value is taken across all the events in the point pattern data with appropriate modification for the edge of the domain. Assuming complete spatial randomness of events on a two-dimensional plane, this simplifies to $K(h) = \pi h^2$ because the expected number of events within radius h of any event is $\lambda\pi h^2$. Simplifying further, $L(h) = (K(h)/\pi)^{0.5} = h$.

$$K(h) = \lambda^{-1} \cdot E[X(h)] \qquad (5.9)$$

The actual L-function, however, will have some random departures from $L(h) = h$ on a finite domain even under complete spatial randomness. To evaluate possible departures from complete spatial randomness, a confidence envelope is constructed. The confidence envelope is the limits of $L(h)$ obtained from a specified number of simulations, each of which has the same number of events scattered randomly across the same domain. If the L-function is below the confidence envelope at any specified distance h, then spatial inhibition (*i.e.* under-dispersion) is suggested. If the L-function is above the confidence envelope, then it suggests non-random clustering. An L-function that diverges from the confidence envelope is indicative of non-homogeneous density across the domain.

Table 5.1: Fisher's index of dispersion tests upon quadrat counts

Image	Number of Quadrats	Mean	Variance	P-value
Top Left	16	8.875	5.717	0.319106
Top Centre	80	1.675	1.792	0.630423
Top Right	16	8.188	10.696	0.376016
Centre Left	80	1.750	2.139	0.174438
Centre	400	0.410	0.393	0.565675
Centre Right	80	1.825	1.691	0.673274
Bottom Left	16	8.813	6.696	0.552068
Bottom Centre	80	1.588	1.486	0.720553
Bottom Right	16	8.125	9.050	0.673274
Composite	9	139.444	125.028	0.963805
	144	8.715	7.632	0.289811

The number of quadrats in each section of the grid depends upon the improved Neubauer grid etched into the counting chamber. The P-value is based upon a two-tailed hypothesis test upon Fisher's index of dispersion (Equation 3.8, Section 3.2.1) with the degrees of freedom equal to one less than the number of quadrats.

Figure 5.3 shows the L-function and associated confidence envelope for the composite data. The L-function suggests slight non-random clustering at a scale of approximately 6 μm because it slightly exceeds the confidence envelope. Of the 1255 microspheres in the domain, visual inspection identified 1205 isolated microspheres, 19 clusters of two microspheres, and 4 clusters of three microspheres. Visual inspection alone cannot determine whether these microspheres are actually clustered or just randomly close together. Clustering is not evident at larger radii because relatively few small clumps were observed; clustering would be more apparent if more of the microspheres were present in clumps and if the clumps were larger.

Based upon the results of this L-function analysis, counts of quadrats that are of a size equivalent to a 6 μm radius may be significantly over-dispersed according to a hypothesis test on Fisher's index of dispersion. To investigate this hypothesis, the `quadratcount` function in the `spatstat` library in R was used to compute Fisher's index of dispersion for counts associated with quadrats of different sizes. The results for the centre-left 1 mm × 1 mm image (the most clumped of the nine images) are shown in Table 5.2. These results show significant non-random dispersion ($P < 0.0115$) for equivalent radii of 5.6-11.3 μm; specifically, they indicate clustering at this scale (because a $P(\chi^2 < D^2)$ near one indicates over-dispersion). Accordingly, a hypothesis test on Fisher's index of dispersion for a particular set of counts is scale-specific and is not a sufficient determination of possible departures from complete spatial randomness at all scales.

In conclusion, point pattern analysis of microspheres using a single sample of stock suspension in a Petroff-Hausser counting chamber (with an enumerated volume of 20 μL) indicated slight clustering, but the departure from the Poisson assumption was only statistically significant (using Fisher's index of dispersion) for small quadrats corresponding to sample volumes of approximately 2-8 nanolitres. In general, the Poisson assumption appears to be valid to describe random sampling error in the Petroff-Hausser chamber, but this result cannot be extrapolated directly to larger sample volumes and may not be representative of the microsphere dispersion in the stock suspension. Point pattern analysis of low numbers of particles in undisturbed large-volume samples is of greatest practical relevance in evaluating the Poisson assumption for random sampling error, but is infeasible.

5.2.3 Validity of the Poisson Assumption in Probabilistic Modelling

Use of the Poisson assumption for random sampling error in the probabilistic models presented herein assumes complete spatial randomness (*i.e.* that the particles are randomly distributed throughout the source according to some locally homogeneous concentration) and that repeated samples are independent. Repeated counts have often been shown to be over-dispersed in relation to the Poisson distribution and it has been alleged that this is due to clumping and may necessitate use of a negative binomial distribution for random sampling error. However, clumping has not been satisfactorily proven on the basis of over-dispersed counts because any variation in concentration, enumerated sample volume, or analytical recovery among the samples will lead to over-dispersed counts even if the particles are randomly distributed. Visual inspection of samples may help to prove clumping, but sample preparation may disrupt the environmental distribution of particles within the sample.

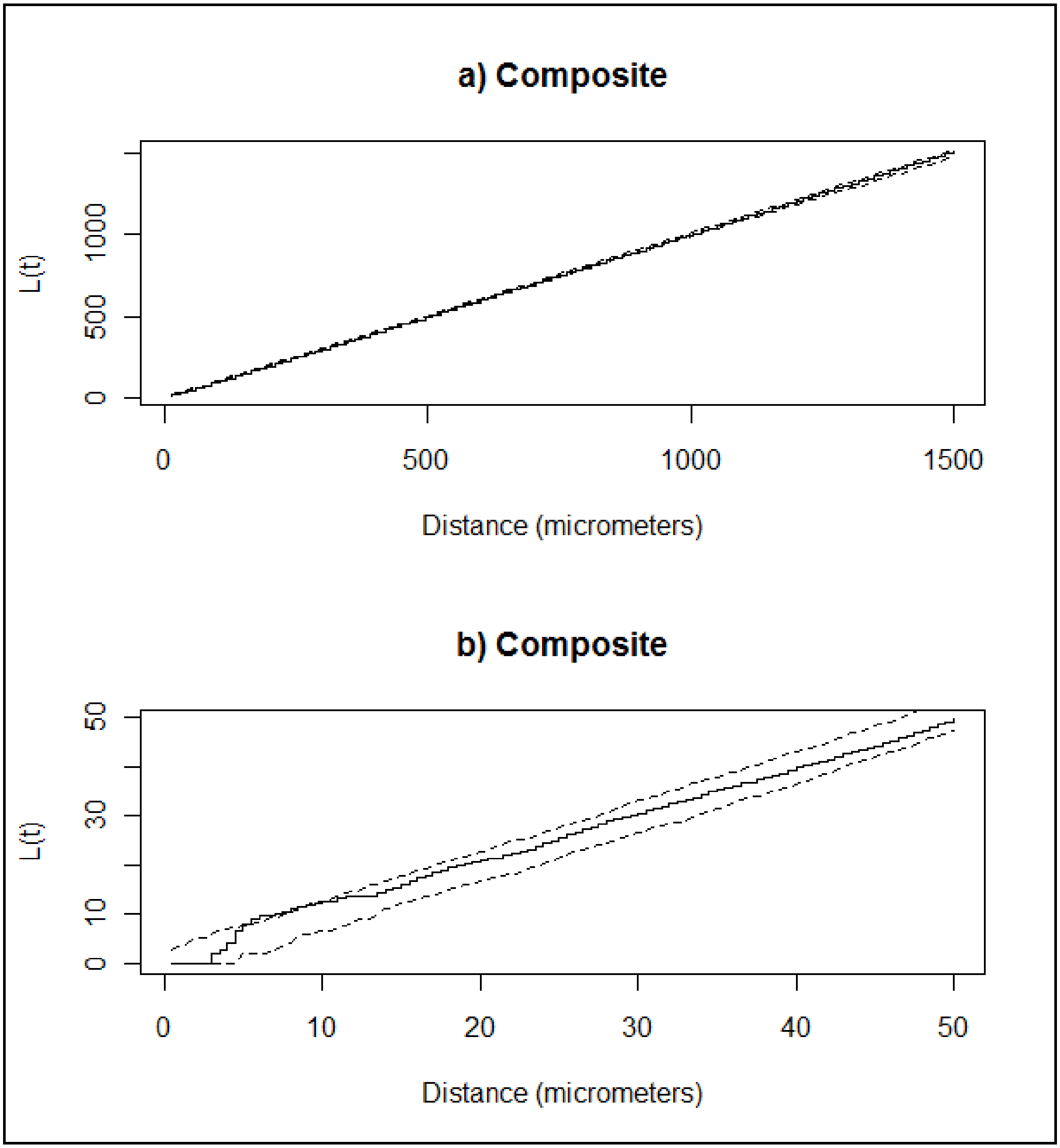


Figure 5.3: Ripley's L-function and confidence envelopes for the composite data

These figures show confidence envelopes (with dashed lines) and the computed L-function (the solid line) for the composite point pattern data. The confidence envelopes and L-function were calculated using the `Kenvl` and `Kfn` functions, respectively, in the `spatial` library in R. Panel (a) shows that the L-function does not diverge from the confidence envelopes (based upon 1000 simulations) as the distance increases. Panel (b) shows the confidence envelopes (based upon 100 simulations) and the L-function at small distances more clearly than panel (a).

Table 5.2: Effect of sample scale upon Fisher's index of dispersion

Grid Partitions	Number of Quadrats	Equivalent Radius (μm)	P(χ^2<D^2)	P-value
2×2	4	282	0.756462	0.487076
5×5	25	113	0.069225	0.138450
10×10	100	56.4	0.235441	0.470882
25×25	625	22.6	0.944868	0.110264
50×50	2500	11.3	0.994290	0.011420 *
80×80	6400	7.1	0.999992	0.000016 *
100×100	10000	5.6	0.998934	0.002132 *
250×250	62500	2.3	0.347692	0.695384

The grid partitions column describes the number of quadrats into which the 1 mm × 1 mm domain was divided horizontally and vertically. The equivalent radius is the radius of a circle with the same area as the square quadrats. The P(χ^2 < D^2) indicates in which portion of the chi-squared distribution the index of dispersion falls: a value near zero suggests under-dispersion while a value near one suggests clustering. The P-values are two-tailed, and values marked with a * indicate significant non-random dispersion at the 5% significance level (all of which suggest clustering).

In concept, alternative probabilistic models to those presented herein can be developed that use different distributions (such as the negative binomial distribution) for random sampling error. Use of the negative binomial model to represent random sampling error, however, is complicated for several reasons.

- It would require case-specific fitted parameters based upon extensive, and often infeasible, replication.
- Any fitted negative binomial model would only be applicable for samples of the same size as the data to which it is fitted (because clustering is scale-specific as demonstrated in Section 5.2.2).
- Counts from a clustered particle suspension may not be negative binomially distributed.

The Poisson distribution is convenient for modelling random sampling error because it does not require any fitted parameters (it depends only upon the unknown concentration and the sample volume). It is stochastically appropriate and can be used for samples of any volume if complete spatial randomness and homogeneous concentration are assumed at the scale of the samples. The negative binomial model may yield a better fit to counts of clustered particles than the Poisson distribution, but it cannot generally be asserted that counts from clustered particle suspensions are negative binomially distributed. Neither model would be appropriate if repeated samples were not independent.

The presented models with Poisson-distributed random sampling error should be used with awareness that the modelled variability in counts (or uncertainty in concentration estimates) is under-predicted if the particles are clustered in the source. Furthermore, if the models address concentration heterogeneity (as discussed in Chapter 6), then the excess variability due to clumping will be attributed to concentration heterogeneity rather than random sampling error. If clustering is severe (*e.g.* patches with very large numbers of particles separated by large volumes that contain few particles), then sampling error may be non-random (*i.e.* it may not be possible to represent the error with a distribution). In such a situation, it may not be possible to make inferences about an average concentration or to assert than any sample is representative of the source as a whole.

5.3 QUANTIFYING UNCERTAINTY IN CONCENTRATION ESTIMATES USING BAYES' THEOREM

Given a probabilistic model that describes the distribution of observable data (*e.g.* particle counts) given an unknown particle concentration, the practical objective is often to evaluate what particle concentration could have yielded specific count data. Bayes' theorem (Section 3.2.2) can be used to obtain a quantitative description of uncertainty in this particle concentration given the probabilistic model (which includes information about the analytical recovery of the enumeration method), the available data, and a prior that describes potentially subjective information about the possible values of concentration. Inference may be based upon a single datum or upon replicates ($i = 1,\ldots,r$). In the beta-Poisson model described in Section 5.1.1, for example, the concentration c is the unknown parameter of interest, the number of particles actually collected in each of the samples ($\{n_i\}$) and the probability of recovery associated with each of the samples ($\{p_i\}$) are unknown nuisance parameters, the beta distribution parameters a,b describing non-constant analytical recovery as well as the sample volumes ($\{V_i\}$) are parameters that are assumed to have precisely known values, and the data are the counts that are obtained ($\{x_i\}$). The posterior for this model is represented by Equation 5.10, in which $g(.)$ is the prior and $h(.)$ is the joint probability of the data given specific values of all unknown parameters calculated using the beta-Poisson model. This proportionality can be simplified to Equation 5.11 because the parameters ($\{V_i\}$,a,b) are precisely known and can be regarded as constants, and $h(.)$ can be rewritten as the likelihood function $L(c,\{n_i\},\{p_i\}|\{x_i\})$.

$$f(c,\{n_i\},\{p_i\}|\{x_i\},\{V_i\},a,b) \propto g(c|\{V_i\},a,b)\cdot h(\{x_i\},\{n_i\},\{p_i\}|c,\{V_i\},a,b) \qquad (5.10)$$

$$f(c,\{n_i\},\{p_i\}|\{x_i\}) \propto g(c)\cdot L(c,\{n_i\},\{p_i\}|\{x_i\}) \qquad (5.11)$$

If the posterior is proper, then the posterior probability density can be evaluated by integration using Equation 5.12. An improper posterior will result when the integral in the denominator of Equation 5.12 is infinite, in which case the posterior is not a distribution. An improper posterior can be used to quantitatively describe uncertainty in unknown parameters (and the relative posterior density of alternative parameter values), but statements of probability cannot be used to describe the uncertainty and the posterior cannot be used in probabilistic models. Proper posteriors are assumed throughout this chapter, and will result from all proper priors and some improper priors. In the present application, the concentration c is of interest and the nuisance parameters $\{n_i\}$,$\{p_i\}$ are not; therefore, the objective is to determine the marginal posterior $f(c|\{x_i\})$. If the likelihood $L(c|\{x_i\})$ can be evaluated (by summing out all possible values of $\{n_i\}$ and integrating out all possible values of $\{p_i\}$), then the marginal posterior density of concentration is represented by Equation 5.13 and (if the posterior is proper) the marginal posterior probability density of concentration is represented by Equation 5.14.

$$f(c,\{n_i\},\{p_i\}\mid\{x_i\})=\frac{g(c)\cdot L(c,\{n_i\},\{p_i\}\mid\{x_i\})}{\int_0^\infty g(c)\cdot L(c,\{n_i\},\{p_i\}\mid\{x_i\})\cdot dc} \tag{5.12}$$

$$f(c\mid\{x_i\})\propto g(c)\cdot L(c\mid\{x_i\}) \tag{5.13}$$

$$f(c\mid\{x_i\})=\frac{g(c)\cdot L(c\mid\{x_i\})}{\int_0^\infty g(c)\cdot L(c\mid\{x_i\})\cdot dc} \tag{5.14}$$

As demonstrated in the preceding equations, the marginal posterior of concentration associated with the beta-Poisson enumeration model (or similarly the negative binomial enumeration model) can be obtained by integration. Explicit integration, however, is often not possible in Bayesian models and such integrals must be numerically approximated. Numerical integration in large Bayesian models, however, can easily become computationally intractable. Accordingly, Markov Chain Monte Carlo methods such as Gibbs sampling are often favoured to obtain approximate posterior probability densities for proper posteriors. Evaluation of the marginal posterior probability density of concentration using integration and Gibbs sampling is discussed in Sections 5.3.1 and 5.3.2 respectively. Analysis of a sample dataset that is used throughout this chapter (and in Emelko *et al.*, 2010a) using numerical integration and Gibbs sampling is presented in Section 5.3.3. Incorporation of internal seed recovery estimates into concentration data analysis is discussed in Section 5.3.4. The beta-Poisson and negative binomial models presented in Section 5.1 both depend upon a distribution for non-constant analytical recovery with the parameters a,b or α,β. Discussion of how to obtain point estimates of these parameters from recovery experiments is addressed in Section 4.4.1. These parameters, however, are uncertain (*e.g.* as discussed in Section 4.4.3), and the Gibbs sampling approach discussed in Section 5.3.2 is expanded to include this uncertainty in Section 5.3.5.

5.3.1 Obtaining the Posterior Concentration Distribution by Numerical Integration

The approach used to obtain posterior concentration distributions by numerical integration using both the beta-Poisson and negative binomial enumeration models was discussed by Emelko *et al.* (2010a). Evaluating the posterior concentration distribution by integration requires that the likelihood $L(c|\{x_i\})$ can be calculated (either explicitly or by numerical approximation) and that the normalizing constant in Bayes' theorem (*e.g.* the numerator in Equation 5.14) can be calculated (by numerical integration in the examples presented herein). Assuming that all samples regarded as replicates are independent and representative of the same source concentration, the likelihood function $L(c|\{x_i\})$ is related to the probability $f(x|c)$ because $L(c|\{x_i\}) = \Pi\, f(x_i|c)$. To evaluate this probability, the nuisance parameters must be integrated out of the probabilistic model. In the beta-Poisson model, n must be summed, and p must be integrated, out of Equation 5.1. It has been shown in Equation 5.3 that summing n out and then integrating p out results in an integral that cannot be solved explicitly. Similarly, Equation 5.15 shows that integrating p out and then summing n out results in a summation that cannot be solved explicitly. The equation for the marginal beta-Poisson count distribution represented by Equation 5.3 is used because the integral is proper and easily approximated, while the summation in Equation 5.15

may converge slowly. An approximation based upon dividing the interval (0,1) into 1000 segments was presented in Emelko *et al.* (2010b), but this approximation is insufficiently accurate when the integrand has a narrow peak. A more efficient algorithm used to approximate the integral in Equation 5.3 is discussed in Appendix C. Unlike the beta-Poisson enumeration model, the marginal count distribution for the negative binomial model (which is obtained by integrating out *p*) can be evaluated explicitly as stated in Equation 5.5.

$$f(x \mid c, V, a, b) = \frac{e^{-cV}}{x!} \cdot \frac{\Gamma(a+b)\Gamma(x+a)}{\Gamma(a)\Gamma(b)} \cdot \sum_{n=x}^{\infty} \frac{(cV)^n \Gamma(n-x+b)}{(n-x)!\Gamma(n+a+b)} \tag{5.15}$$

The posterior density functions for concentration associated with the beta-Poisson and negative binomial enumeration models are presented in Equations 5.16 and 5.17, in which *g*(*c*) represents the prior. The posterior probability density function (assuming that the posterior is proper) can be approximated by numerical integration of the posterior density across an interval (0,c_{max}) outside of which the posterior density is very small. If increases in c_{max} or the number of segments in the numerical integration do not substantially refine the approximated integral, then the numerical integration has converged and the posterior probability density associated with any concentration value can be calculated by dividing its associated posterior density by the value of the integral. Assuming improper uniform priors on the parameter space $c > 0$ (as per Emelko *et al.*, 2010a), the posterior probability density function for concentration is directly proportional to the likelihood function. Further research is needed to establish whether or not the posterior associated with an improper uniform prior yields a proper posterior (*i.e.* whether or not the integral converges to a finite value) for all possible parameter values. Preliminary results have shown that the likelihood does converge upon zero as concentration increases, but may do so very slowly if the non-constant analytical recovery distribution includes many recovery values near zero. If the integral does not converge, then the posterior is not a distribution and uncertainty in the particle concentration cannot be described using probabilities. Use of a weakly informative proper prior (when the prior is intended to represent ignorance about the possible parameter values) or an informative proper prior based upon previous data or the analyst's subjective beliefs will yield a proper posterior. The numerical integration approach is used to evaluate uncertainty in the particle concentration estimated from replicate enumeration data in Section 5.3.3. Sample code is provided in Appendix C.

$$f(c \mid \{x_i\}) \propto g(c) \cdot \prod_{i=1}^{r} c^{x_i} \cdot \int_0^1 e^{-cV_i p} p^{x_i+a-1} (1-p)^{b-1} \, dp \tag{5.16}$$

$$f(c \mid \{x_i\}) \propto g(c) \cdot \prod_{i=1}^{r} \frac{c^{x_i}}{(cV_i\beta+1)^{x_i+\alpha}} \tag{5.17}$$

5.3.2 Obtaining the Posterior Concentration Distribution by Gibbs Sampling

Information about the joint posterior distribution of the parameters (*i.e.* concentration as well as the nuisance parameters) can be obtained by Gibbs sampling if the posterior is proper. The Gibbs sampling algorithm for the beta-Poisson enumeration model (based upon Nahrstedt and Gimbel,

1996) was developed by Emelko (2001) and algorithms for both the beta-Poisson and negative binomial enumeration models were presented by Emelko *et al.* (2010a). Development of both algorithms is discussed herein and additional information for these and for the internal seed enumeration model is provided in Appendix B.3. Gibbs sampling (Section 3.2.2) can be used to generate a Markov chain of values that are collectively representative of a posterior distribution (assuming convergence). Use of Gibbs sampling for this purpose requires (1) application of Bayes' theorem with an appropriate prior for concentration to obtain a joint posterior density function for all of the unknown parameters in the model, (2) derivation of the conditional posterior density function for each unknown parameter in the model (including the nuisance parameters), (3) preparation of algorithms to draw random numbers from each of the conditional posterior distributions, and (4) specification of initial values for each of the unknown parameters. When an improper prior is used, evaluation of whether or not the posterior is proper should precede Gibbs sampling (Hobert and Casella, 1996) because the Gibbs sampling results do not always clearly indicate impropriety.

Using $g(c)$ to represent the prior for concentration, application of Bayes' theorem to the beta-Poisson enumeration model (Equation 5.1) and the negative binomial enumeration model (Equation 5.4) yields the joint posterior density functions represented by Equations 5.18 and 5.19 respectively. The nuisance parameters are retained in these models (rather than being integrated out) because the resulting Gibbs sampling algorithm is easier. If, for example, the rates of recovery ($\{p_i\}$) were integrated out of the negative binomial model, then the Gibbs sampling algorithm would be sampling directly from the posterior concentration distribution (which is much more difficult than the algorithm developed below). The prior that is required to use Bayes' theorem only addresses concentration; prior information about the nuisance parameters must be incorporated into the probabilistic model that is used to derive the likelihood function.

$$f\left(c,\{n_i\},\{p_i\}\mid\{x_i\}\right)\propto \\ g(c)\cdot\prod_{i=1}^{r}\left[\frac{e^{-cV_i}\left(cV_i\right)^{n_i}}{n_i!}\right]\left[\frac{n_i!}{x_i!\left(n_i-x_i\right)!}p_i^{x_i}\left(1-p_i\right)^{n_i-x_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p_i^{a-1}\left(1-p_i\right)^{b-1}\right] \tag{5.18}$$

$$f\left(c,\{p_i\}\mid\{x_i\}\right)\propto g(c)\cdot\prod_{i=1}^{r}\left[\frac{e^{-cV_ip_i}\left(cV_ip_i\right)^{x_i}}{x_i!}\right]\left[\frac{1}{\beta^{\alpha}\Gamma(\alpha)}p_i^{\alpha-1}e^{-p_i/\beta}\right] \tag{5.19}$$

The conditional posterior distribution functions for the unknown parameters in the beta-Poisson enumeration model (c, n_i, and p_i), as derived from Equation 5.18, are summarized in Equations 5.20, 5.21, and 5.22 respectively (with some modifications from Emelko *et al.*, 2010a).

$$Dfc(c)\propto g(c)\cdot e^{-c\sum V_i}c^{\sum n_i} \tag{5.20}$$

$$Dfc(n_i)\propto\frac{\left(cV_i\left(1-p_i\right)\right)^{n_i}}{\left(n_i-x_i\right)!}\rightarrow n_i-x_i\sim POISSON\left(cV_i\left(1-p_i\right)\right) \tag{5.21}$$

$$Dfc(p_i)\propto p_i^{x_i+a-1}\left(1-p_i\right)^{n_i-x_i+b-1}\rightarrow p_i\sim BETA\left(x_i+a,n_i-x_i+b\right) \tag{5.22}$$

Similarly, the conditional posterior distribution functions for c and p_i in the negative binomial enumeration model (as derived from Equation 5.19) are summarized in Equations 5.23 and 5.24 respectively (as per Emelko *et al.*, 2010a with addition of the prior $g(c)$).

$$Dfc(c) \propto g(c) \cdot e^{-c\sum V_i p_i} c^{\sum x_i} \tag{5.23}$$

$$Dfc(p_i) \propto e^{-p_i(cV_i + 1/\beta)} p_i^{x_i+\alpha-1} \rightarrow p_i(cV_i + 1/\beta) \sim GAMMA(x_i + \alpha, 1) \tag{5.24}$$

As demonstrated in Emelko *et al.* (2010a), the conditional posterior distributions for concentration are gamma-distributed if an improper uniform prior on the parameter space $c > 0$ is used: $c\Sigma V_i \sim GAMMA(\Sigma n_i + 1,1)$ for the beta-Poisson model and $c\Sigma V_i p_i \sim GAMMA(\Sigma x_i + 1,1)$ for the negative binomial model. Accordingly, gamma-distributed informative priors for concentration will also yield convenient conditional posteriors. Emelko (2001) used the improper prior $g(c) = c^{-1}$ for the beta-Poisson model because Jeffreys (1961) recommended this prior for continuous parameters on the positive semi-infinite domain. The associated conditional posterior for concentration is $c\Sigma V_i \sim GAMMA(\Sigma n_i,1)$. If $\Sigma n_i = 0$ (which has some non-zero probability if $\Sigma x_i = 0$), then the conditional posterior for concentration is not a distribution and the posterior is improper. Similarly, the resulting conditional posterior for concentration in the negative binomial model, $c\Sigma V_i p_i \sim GAMMA(\Sigma x_i\ ,1)$, indicates an improper posterior if $\Sigma x_i = 0$. One of the advantages of Bayesian analysis is that uncertainty in the concentration can be inferred from individual or replicate non-detect data. Consequently, the prior $g(c) = c^{-1}$ was not used in Emelko *et al.* (2010a) and is not used herein.

Herein, improper uniform priors are used and the Gibbs sampling algorithm is relatively simple because all of the conditional posteriors are related to distributions for which random sampling algorithms are commonly available. In several of the conditional posteriors, the distribution of the parameter is not stated explicitly (*e.g.* the random variable in Equation 5.21 is $n_i – x_i$). The parameter value is obtained by generating a random deviate using the specified distribution and solving for the parameter of interest algebraically (*e.g.* n_i is the sum of x_i and a Poisson-distributed random number in Equation 5.21). The Gibbs sampling algorithms for the beta-Poisson and negative binomial enumeration models are summarized in Figures 5.4 and 5.5. These figures also show how to incorporate uncertainty in the non-constant analytical recovery distribution parameters (a,b or α,β), which is addressed in Section 5.3.5.

Gibbs sampling requires initial values for the unknown parameters. Any value within the feasible parameter space (*i.e.* $c > 0$, $n_i > x_i$, $0 < p_i < 1$) is acceptable, although a shorter burn-in is typically possible if relatively probable parameter values are chosen. In Emelko *et al.* (2010a), the Gibbs sampling sequences were arranged in a different order than what is presented in Figures 5.4 and 5.5. The recommended initial values were $n_i = x_i \cdot (a + b) / a$ (which needs to be rounded to the nearest integer) in the beta-Poisson model and $c = \Sigma x_i / \alpha\beta\Sigma V_i$ in the negative binomial model (which needs to be assigned a positive value when $\Sigma x_i = 0$). Herein, the Gibbs sampling sequences are arranged in a way that allows sample-specific recovery information (*i.e.* a,b or α,β are not necessarily equal for all of the replicate enumeration data). The initial recovery values used herein are $p_i = a_i / (a_i + b_i)$ for the beta-Poisson enumeration model and $p_i = \alpha_i\beta_i$ for the negative binomial enumeration model. The Gibbs sampling algorithm for the beta-Poisson enumeration model that is shown in Figure 5.4 also requires an initial concentration value. The value used herein is $c = \Sigma x_i / \Sigma V_i p_i$ unless $\Sigma x_i = 0$, in which case Σx_i is assigned a value of 0.5. Gibbs sampling is used to evaluate uncertainty in the particle concentration estimated from replicate data in Section 5.3.3. Sample code is provided in Appendix C.

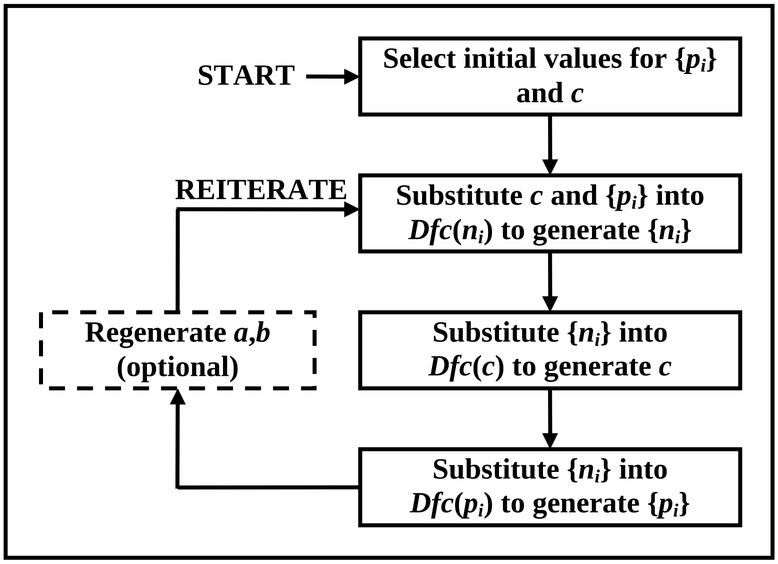

START
Select initial values for {p_i} and c
REITERATE
Substitute c and {p_i} into $Dfc(n_i)$ to generate {n_i}
Regenerate a,b (optional)
Substitute {n_i} into $Dfc(c)$ to generate c
Substitute {n_i} into $Dfc(p_i)$ to generate {p_i}

Figure 5.4: Gibbs sampling algorithm for the beta-Poisson enumeration model
The Gibbs sampling algorithm begins with a burn-in and ends after a user-specified number of iterations. Uncertainty in the parameters of the beta distribution describing non-constant analytical recovery can be addressed by using new estimated parameter values from a posterior distribution (*e.g.* using Gibbs sampling as discussed in Section 4.4.3).

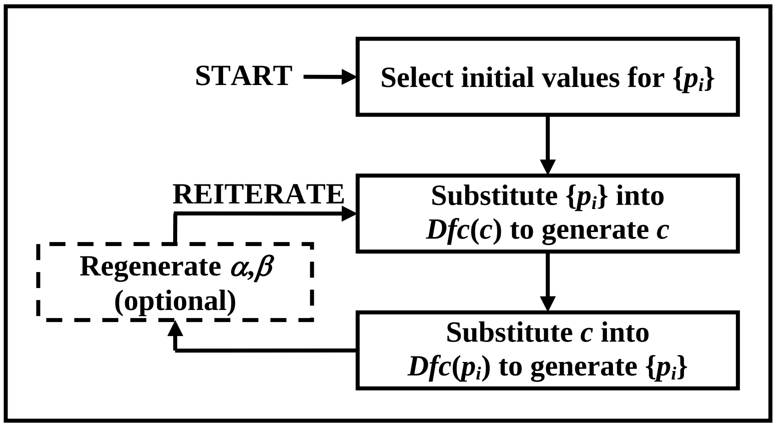

START
Select initial values for {p_i}
REITERATE
Substitute {p_i} into $Dfc(c)$ to generate c
Regenerate α,β (optional)
Substitute c into $Dfc(p_i)$ to generate {p_i}

Figure 5.5: Gibbs sampling algorithm for the negative binomial enumeration model
The Gibbs sampling algorithm begins with a burn-in and ends after a user-specified number of iterations. Uncertainty in the parameters of the gamma distribution describing non-constant analytical recovery can be addressed by using new estimated parameter values from a posterior distribution (*e.g.* using Gibbs sampling as discussed in Appendix B.1.3).

5.3.3 Bayesian Quantification of Uncertainty in Example Concentration Estimates

The enumeration data shown in Table 5.3 are used herein to exemplify use of the numerical integration and Gibbs sampling approaches presented in the preceding sections. As described in Emelko *et al.* (2010a) these data represent an experiment in which the $\log_{10}$-reduction of a treatment technology is being investigated (see Section 5.4 for further discussion of these experiments) and "were simulated using the beta-Poisson model with initial and final concentrations of 50 and 0.5 microorganisms/L, respectively, and recovery parameters $a = 287.08$ and $b = 94.76$ ($\mu = 75.18\%$, $\sigma = 2.208\%$)".

Table 5.3: Example enumeration data (Emelko *et al.*, 2010b)

	Initial		**Final**			
Volume (L)	10	10	50	50	50	50
Count	376	388	16	16	19	29

Results based on numerical integration and Gibbs sampling are plotted in Figure 5.6. In the numerical integration approach, the step-size and upper limit for the initial and final concentration estimates are (0.05, 63.4) and (0.0005, 0.97) respectively. The Gibbs sampling results are based on 30,000 iterations following a burn-in of 1000 iterations. The two left panels in Figure 5.6 show the posterior density functions estimated by numerical integration as well as 95% credible intervals. The intervals shown are the narrowest 95% credible intervals because they contain only the most probable concentration values with cumulative posterior probability of 95%. Equal-tailed 95% credible intervals would not have the same posterior probability density at either limit unless the distribution was symmetrical and would be wider than the intervals shown here (because the interval would include some concentration values in one tail of the distribution that have lower posterior probability density than some values that are excluded in the other tail). The credible intervals for initial concentration (46.78, 55.10) and final concentration (0.423, 0.659) include the true concentration values of 50 and 0.5 microorganisms/L, respectively, with which the enumeration data were simulated. Given a model that describes the random errors in the enumeration data appropriately and a suitable prior, the 95% credible interval will contain the true value of the concentration with 95% probability. The two right panels in Figure 5.6 show the cumulative posterior probability functions obtained by numerical integration and by Gibbs sampling. The curves coincide for both the initial and final concentrations, which suggests that these particular Gibbs sampling sequences have converged after 30,000 iterations (following a burn-in of 1000 iterations).

Mixing and convergence are important properties of Gibbs sampling methods because a Markov chain with poor mixing will require a large burn-in and many iterations to reach convergence and because a Markov chain that has not converged will not yield parameter values that are collectively representative of the posterior. Figure 5.7a shows the mixing of the Gibbs sampling sequences for both concentration datasets using the recommended initial parameter values. These chains both

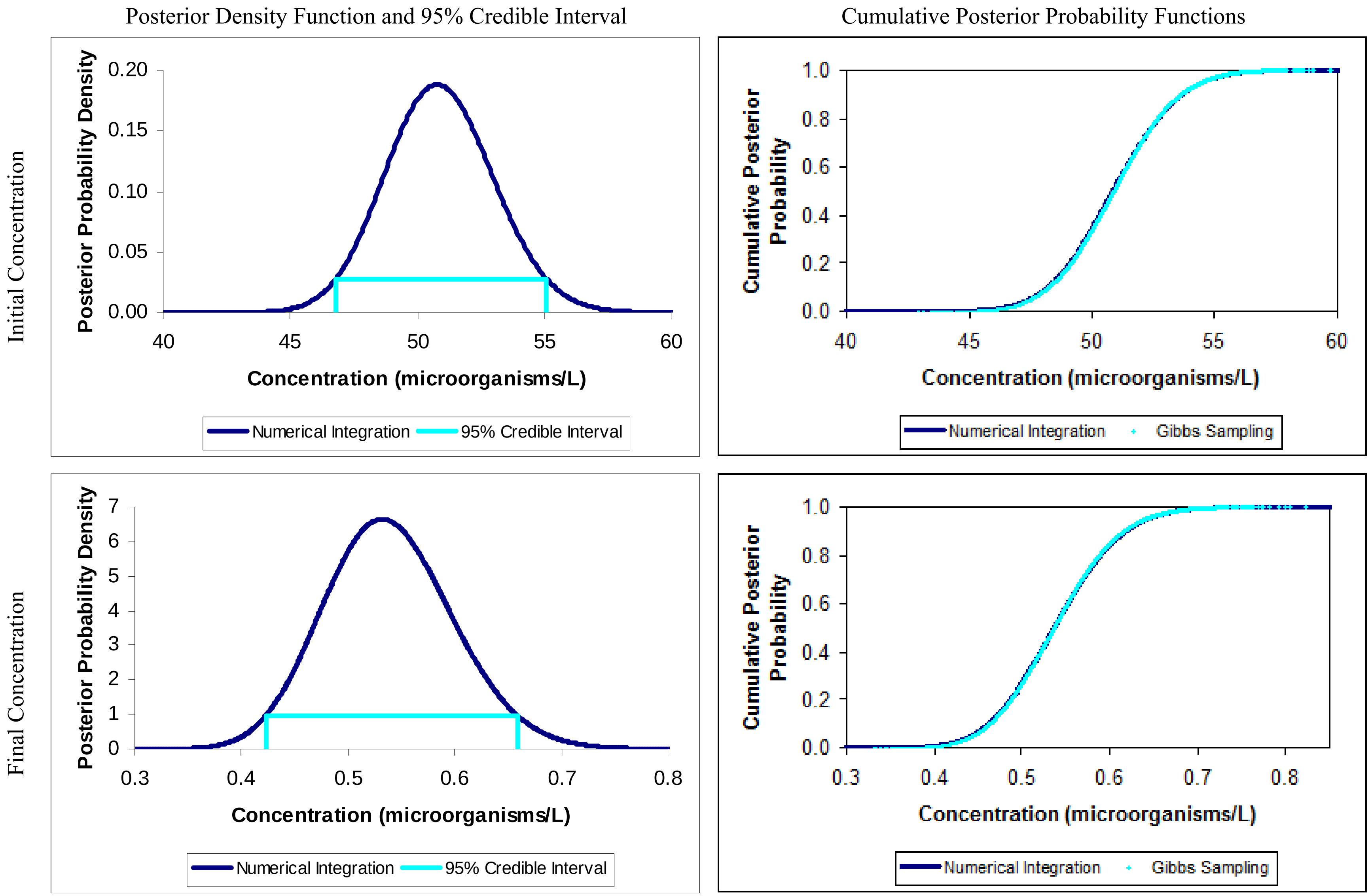


Figure 5.6: Posterior concentration distributions associated with the Table 5.3 data

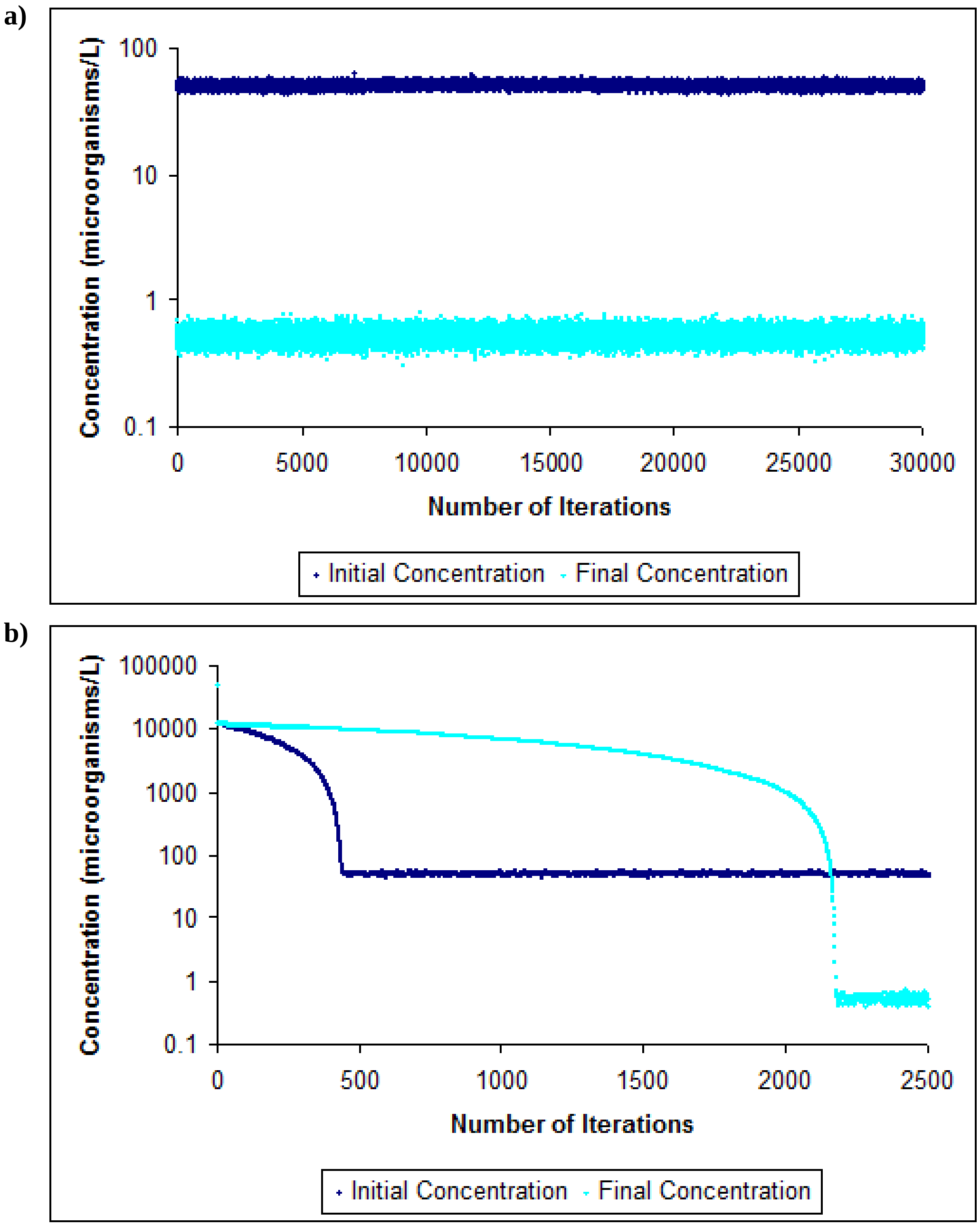


Figure 5.7: Gibbs sampling sequences used to evaluate mixing and convergence

These Gibbs sampling results are based on analysis of the Table 5.3 enumeration data using the beta-Poisson enumeration model with an improper uniform prior for concentration. The initial concentration values were (a) the recommended concentration estimates described in Section 5.3.2 and (b) 50,000 microorganisms/L.

exhibit exceptionally good mixing and show that a small burn-in (certainly less than the 1000 iterations that are used in this example) is acceptable for the selected initial parameter values. Figure 5.7b shows the Gibbs sampling sequences obtained when the initial concentration values are both set to the improbably high value of 50,000 microorganisms/L. In both sequences, the concentration values fall to typical values within 2500 iterations. These results suggest that the posterior is proper because an improper posterior associated with a prior that does not converge to zero for infinite concentration values would likely result in a Gibbs sampling sequence in which concentration diverges to high values. The Gibbs sampling sequences were observed to reach typical concentration values after as few as 5 iterations when improbably low initial concentration values (*e.g.* 10^{-250} microorganisms/L) were used.

The Table 5.3 data were also analyzed using the negative binomial enumeration model and the gamma distribution parameters $\alpha = 1236.16$ and $\beta = 0.000608$ (which correspond to $\mu = 75.18\%$ and $\sigma = 2.138\%$). The resulting posterior distributions (not shown) were very similar to those shown in Figure 5.6 (because the recovery distribution has a similar mean and standard deviation and only a slightly different shape). Analysis of mixing and convergence for this scenario (not shown) indicated even better mixing than was demonstrated in Figure 5.7a and more rapid convergence from improbable initial parameter values[1] than demonstrated in Figure 5.7b.

The Bayesian approaches to quantify uncertainty in concentration estimates that are used herein are not limited to replicate data or to positive counts. For example, the same approaches can be used to evaluate uncertainty in a single non-detect sample (which would have an associated unbiased concentration estimate of 0 microorganisms/L). Figure 5.8 presents the results of analysis upon a single non-detect 1 L sample using Gibbs sampling. The sequence of 31,000 concentration values generated using the beta-Poisson enumeration model and the parameters $a = 3.34$, $b = 32.90$ (based upon the case study data in Section 4.4.4) is shown in Figure 5.8a. Similarly, the sequence of 31,000 concentration values generated using the negative binomial enumeration model and the parameters $\alpha = 3.781$, $\beta = 0.0244$ is shown in Figure 5.8b. These non-constant analytical recovery distributions have a low mean (9.22%), relatively high standard deviation (4.74%), and can yield very low recovery values (the probability that recovery is less than 1% is 0.0023 and 0.0014 for the two models, respectively). Figure 5.8a demonstrates somewhat poor mixing, particularly at high concentrations, while Figure 5.8b shows good mixing. As described in Emelko *et al.* (2010b), wide posterior distributions result when the non-constant analytical recovery distribution includes many recovery values near zero (because very high concentrations can yield low counts when recovery is near zero), and these tend to result in slow-mixing Gibbs sampling sequences. The highest generated concentration values are 732 and 1366 microorganisms/L for the two models, respectively. Gibbs sampling will work with any data and parameter values, assuming that the posterior is proper, but may require a longer burn-in and more iterations to properly characterize the posterior when mixing is

[1] The Gibbs sampling algorithm used here for the negative binomial enumeration model is different than shown in Figure 5.5 because it starts with a specified concentration.

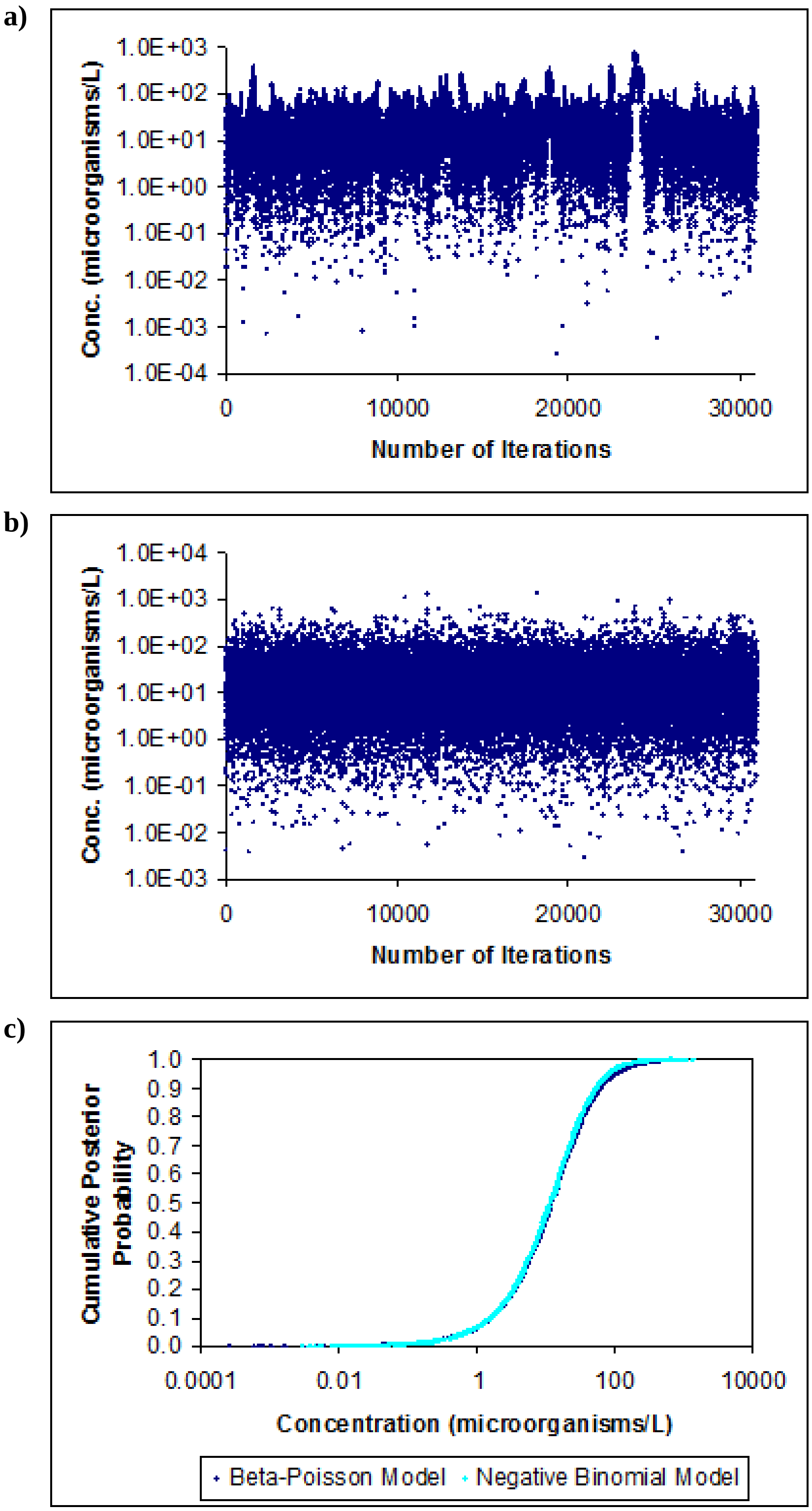


Figure 5.8: Gibbs sampling sequences with non-detect data and low recovery

The sequences of 31,000 iterations (of which the first 1000 were regarded as burn-in) associated with a one-litre non-detect sample are shown using (a) beta-distributed non-constant analytical recovery and (b) gamma-distributed non-constant analytical recovery (both with μ = 9.22% and σ = 4.74%). Panel (c) shows the associated cumulative posterior probability functions.

poor. The cumulative posterior probability functions shown in Figure 5.8c (based upon 30,000 iterations following a burn-in of 1000 iterations) are similar because beta and gamma distributions with the same mean and standard deviation have similar shape when the probability that recovery exceeds 100% in the gamma model (10^{-14} in the present example) is trivially small.

Use of the numerical integration approach for this example is also somewhat complicated because the posterior is very wide (*i.e.* the posterior density converges slowly upon zero). A very small step-size is needed to properly evaluate the posterior density of concentrations near zero, but the numerical integration must also proceed to high concentration values (*e.g.* 16,000 microorganisms/L) for the integral to converge.

5.3.4 Bayesian Analysis of Datasets with Internal Seed Recovery Data

In the preceding application of Bayes' theorem to obtain posterior distributions for concentration using the beta-Poisson enumeration model, it was assumed that the information about analytical recovery is provided in the form of a beta-distributed non-constant analytical recovery distribution that was estimated from an independently conducted recovery experiment. When analytical recovery is addressed using internal seed recovery estimates, so that some or all of the counts are accompanied with sample-specific recovery information, a somewhat different model is obtained (*e.g.* Section 5.1.3). When Bayes' theorem is applied to the probabilistic model represented by Equation 5.8 (with single or replicate data, each of which is accompanied by internal seed recovery data), each sample's analytical recovery is an unknown parameter of interest for which a prior must be specified. In contrast, analytical recovery is a nuisance parameter in the beta-Poisson model because prior information about its distribution is included in the probabilistic model in the beta distribution for non-constant analytical recovery. Application of Bayes' theorem to a model for single or replicate concentration data, each of which is accompanied with internal seed recovery data, is addressed in Appendix B.3.2.

Rather than using a separate probabilistic model for samples with internal seed recovery data, it is possible to incorporate internal seed recovery data directly into the beta-Poisson model as discussed in Emelko *et al.* (2010b) and Schmidt and Emelko (2010a). In the beta-Poisson model, the beta distribution can be regarded as a distribution representing non-constant analytical recovery or (in a Bayesian context) as an informative prior about the possible values of analytical recovery in each sample. Accordingly, internal seed recovery data can be used to construct an informative prior for each sample's analytical recovery, and this can be incorporated directly into Bayesian analysis of the beta-Poisson model (in the form of sample-specific beta distribution parameters a_i,b_i) if these informative priors are beta-distributed. This situation is particularly convenient if some replicate enumeration data are paired with internal seed recovery data and others are not (because the varying information can easily be incorporated as sample-specific beta distribution parameters). Otherwise a somewhat more complicated probabilistic model with two different types of recovery information (and associated distributions and parameters) would need to be constructed.

In Section 5.1.3, it is shown that the number of internally seeded particles (x^*) that are observed in a sample into which n^* particles were seeded and that has analytical recovery p can be modelled by a binomial distribution: $x^* \sim BINOMIAL(n^*,p)$. Application of Bayes' theorem to this model yields the posterior for analytical recovery $f(p) \propto g(p) \cdot p^{x^*} (1-p)^{n^*-x^*}$. If the prior $g(p)$ is beta-distributed (the conjugate prior to the binomial distribution as described in Section 3.2.2), then the posterior is also beta-distributed. If a beta-distributed prior with parameters a^*,b^* is used, then the posterior has parameters $a = x^* + a^*$, $b = n^* - x^* + b^*$. The beta-distributed prior can be an informative prior representing non-constant analytical recovery or it can be a relatively uninformative uniform prior (a special case of the beta distribution with $a^* = 1$, $b^* = 1$). Internal seed recovery data are used in example Bayesian concentration analyses in Section 6.3.2.

5.3.5 Including Uncertainty in Recovery Distribution Parameters in Gibbs Sampling

The preceding analyses of enumeration data based upon the beta-Poisson and negative binomial models assumed precisely known values of the parameters (a,b or α,β) in the distributions representing non-constant analytical recovery. The supplied parameter values are described in Emelko *et al.* (2010a) as estimates obtained from the results of an independently conducted recovery experiment (which must be representative of the sample composition, methodology, and laboratory). Appropriate point estimates should be obtained using maximum likelihood estimation and a probabilistic model that suitably describes the seeding and analytical error in the recovery experiment (Schmidt *et al.*, 2010a; Section 4.4.1). Even if an appropriate point estimation method is used, there will still be uncertainty in the parameter values (*e.g.* Section 4.4.3). Accordingly, a more rigorous analysis must also account for the uncertainty in the parameters of the non-constant analytical recovery distribution. Petterson *et al.* (2007) used a Bayesian Markov Chain Monte Carlo approach to evaluate uncertainty in the parameters of the beta distribution describing non-constant analytical recovery, but integrated this information into analysis of enumeration data using a non-Bayesian Monte Carlo approach (as discussed in Schmidt and Emelko, 2010b and in Section 6.1.2).

As illustrated in Figures 5.4 and 5.5, uncertainty in the non-constant analytical recovery distribution parameters can easily be incorporated into the Gibbs sampling algorithm by supplying a new pair of parameter values (a random sample from their joint posterior distribution) for each iteration. This approach essentially integrates the posterior distribution describing uncertainty in these parameters into the analysis using Monte Carlo (because the distribution describing their uncertainty is independent of the Bayesian model into which the parameter values are being inserted). This approach is used because the information about non-constant analytical recovery results strictly from carefully controlled recovery experiments and an appropriate prior on the parameters. In contrast, a fully Bayesian model would treat the non-constant analytical recovery distribution parameters as unknown parameters of interest (*i.e.* a prior would need to be provided for them and they would be included in the posterior distribution). Such a model would either address the enumeration data and recovery experiment data concurrently (perhaps with a relatively uninformative prior upon the recovery distribution parameters) or would use the separate Bayesian analysis of the recovery experiment data as an informative prior upon the recovery distribution parameters. In either case, the posterior for these parameters would include some additional information about analytical recovery from the enumeration dataset, but would be affected by the choice of prior on concentration and would also be affected if samples representing different concentrations are misclassified as replicates. It is, therefore, preferable to not include these parameters as unknown parameters of interest in the Bayesian analysis. Uncertainty in the non-constant analytical recovery parameters is not addressed in the numerical integration approach because the posterior distribution for these parameters generally cannot be explicitly determined and because the resulting integrations would be intractable.

To properly integrate uncertainty in the non-constant analytical recovery distribution parameters into the Gibbs sampling algorithm using Monte Carlo, the supplied parameter pairs should be drawn randomly from their joint posterior distribution. Assuming that an appropriate probabilistic model has been chosen to analyze the recovery data and that a suitable prior that results in a proper posterior has been chosen, the Gibbs sampling process discussed in Section 4.4.3 can be used to generate pairs of the non-constant analytical recovery distribution parameters. The order of a reasonably converged sequence of parameter pairs can be randomized so that possible poor mixing in the Gibbs sampling analysis of the recovery data does not compromise the mixing of the Gibbs sampling analysis of the concentration data. If the former Gibbs sampling process has not converged, then the resulting parameter pairs will not be optimally representative of their associated joint posterior distribution; however, even imperfectly converged results will often reflect much of the uncertainty in these parameters. In the interest of expedience, for example, it may be possible to provide an acceptable consideration of uncertainty in these parameters by generating a set of just a few thousand pairs and then drawing these pairs from the set randomly (with replacement so that pairs can be used more than once). If the posterior concentration distribution is not very sensitive to uncertainty in the non-constant analytical recovery distribution parameters, then this source of uncertainty can be omitted from the analysis.

To illustrate the effects of uncertainty in the non-constant analytical recovery parameters upon uncertainty in concentration estimates, the Table 5.3 data are analysed again with pairs of beta distribution parameters that are representative of their posterior distribution. The Table 5.3 enumeration data were simulated using the beta-Poisson enumeration model with a = 287.08, b = 94.76 (the maximum likelihood estimates associated with the beta-Poisson recovery model for the Table 4.4 recovery data). Accordingly, uncertainty in the beta distribution parameters is addressed herein by using parameter pairs that were generated by Gibbs sampling using the beta-Poisson recovery model and the Table 4.4 data. The Gibbs sampling process for the beta-Poisson recovery model is similar to that of the beta-binomial recovery model (Section 4.4.3) except that the actual seed doses ($\{n_i\}$) are related by a Poisson distribution to the presumed seed doses ($\{\lambda_i\}$), and that there is an extra conditional posterior distribution to build into the Gibbs sampling algorithm (Appendix B.1.2). A total of 26,000 parameter pairs were generated using the prior $g(a,b) = (ab)^{-1}$ and a burn-in of 1000 iterations, and storing every tenth parameter pair after the burn-in. The Gibbs sampling sequence was found to mix slowly (as is typical when high parameter values are common), so the resulting sequence may not be a fully converged representation of the joint posterior distribution of the beta distribution parameters. Nonetheless, it is assumed that the sequence provides an adequate representation of the uncertainty in these parameters. The generated parameter pairs were integrated into the beta-Poisson enumeration model Gibbs sampling algorithm illustrated in Figure 5.4. 25,000 generated concentration values (following a burn-in of 1000 iterations) were used to represent the posterior concentration distribution. This analysis was compared to Gibbs sampling using the beta-Poisson recovery model maximum likelihood parameter estimates (a = 287.08, b = 94.76). For both the initial and final concentration data, the resulting cumulative posterior probability functions (not shown) coincide. In this particular scenario, therefore, accounting for uncertainty in the beta distribution parameters is no more informative about uncertainty in concentration than simply using the maximum likelihood parameter estimates.

In general, uncertainty in the non-constant analytical recovery distribution parameters relates partially to uncertainty in the mean analytical recovery and mostly to uncertainty in the standard deviation of analytical recovery (because second moments, such as the variance, are more uncertain than first moments such as the mean). Therefore, addressing uncertainty in the non-constant analytical recovery distribution parameters is most impactful (in terms of refining posterior concentration distributions) in scenarios in which the uncertainty in the concentration estimate is sensitive to the variance of analytical recovery. As will be discussed in Sections 5.6 and 8.2.2, the uncertainty in concentration estimates is most sensitive to the variability of analytical recovery when relatively high counts are obtained (because the variability in repeated enumerations depends upon variable analytical recovery more than random sampling error). Addressing this uncertainty will also be most impactful when the posterior predictive distribution of analytical recovery is substantially different from the recovery distribution associated with the maximum likelihood estimates; otherwise, the uncertainty in the non-constant analytical recovery distribution parameters would have little impact upon concentration estimates because it would have little impact upon analytical recovery itself.

5.4 QUANTIFYING UNCERTAINTY IN LOG-REDUCTION ESTIMATES

The capacity of a treatment process (*e.g.* filtration, disinfection) to reduce the concentration of microorganisms (or discrete particles used as surrogates for microorganism removal) is often of interest. This capacity is generally described as log-reduction (or log-removal for a physico-chemical removal process such as filtration). Given an initial concentration c_1 and a final concentration c_2, the log-reduction is calculated as $\rho = \log_{10}(c_1/c_2)$. In general, c_1 and c_2 are uncertain due to inevitable measurement errors (although c_1 may be precisely known if it is a carefully seeded concentration) and log-reduction will therefore also be uncertain. If c_1 and c_2 are both highly uncertain, then extreme log-reduction values can easily result from an over-estimate of c_1 and an under-estimate of c_2 or vice versa. Accordingly, it is imperative to consider the uncertainty in log-reduction estimates and the concentration values upon which they are based.

The objective in this section is to quantitatively describe uncertainty in log-reduction using probabilistic models. Section 5.4.1 describes ‘probability of passage’ models that have been used by Teunis *et al.* (1997; 1999; 2009), Teunis and Havelaar (1999; 2002), and Medema *et al.* (2003). A partially Bayesian method that yields a distribution describing uncertainty in a single log-reduction estimate based on posterior distributions for the initial and final concentrations (Emelko, 2001; Emelko *et al.*, 2008; 2010a) is discussed in Section 5.4.2. A fully Bayesian approach to generate a posterior distribution for a single log-reduction estimate is presented in Section 5.4.3.

5.4.1 'Probability of Passage' Models

The basic premise of a 'probability of passage' model is that the number of microorganisms (or discrete particles) in a specific volume of water following a treatment process is reduced from the number initially present by an unknown probability of passage (that is equal for all of the target microorganisms in the sample). This section summarizes 'probability of passage' models that are used in various applications and how these types of models have been used to evaluate log-reduction.

Dose-response and analytical recovery 'probability of passage' models

The 'probability of passage' concept is applied in the widely used exponential and beta-Poisson dose-response models (*e.g.* Haas, 1983) and in recovery models that regard analytical recovery as the probability that each particle present in the sample will be observed (the similarity between the two types of models is also discussed in Section 4.2.8). In the exponential and beta-Poisson dose-response models, it is assumed that a consumer will imbibe a volume of drinking water containing some specific number of a particular type of pathogen. The number of such pathogens in the imbibed volume is random and is related to the dose (the product of source concentration and imbibed volume) using a Poisson distribution (similar to the Poisson model for random sampling error used in Section 4.2.3). It is then assumed that there is some host-pathogen interaction that will inactivate some number of the imbibed pathogens; each pathogen is assumed to have an equal probability of surviving this interaction and successfully initiating an infection. The number of the imbibed pathogens that initiate the infection is implicitly modelled using a binomial distribution (similar to the binomial model for analytical error used in Sections 4.2.2 and 4.2.3). Together, these two models yield a Poisson distribution with the mean equal to the product of dose and the probability that each consumed pathogen will survive to initiate an infection in the host. The exponential dose-response model computes the probability that at least one of the consumed pathogens in a specific volume of water will survive to initiate infection. The beta-Poisson dose response model is similar, except that the host-response interaction (the probability that a consumed pathogen will survive to initiate infection) is assumed to vary according to a beta distribution throughout a population of consumers (similar to the beta-distributed non-constant analytical recovery used in Sections 4.2.2 and 4.2.3).

In dose-response models, the 'probability of passage' is the probability that a consumed pathogen will survive to initiate infection. Analytical recovery in the beta-binomial and beta-Poisson recovery models can also be regarded as a probability of passage. The main difference between these two types of 'probability of passage' models is that the number of imbibed pathogens that initiated the infection observed in a particular host is immeasurable. The dose-response models, therefore, are based upon the frequency of detected infections for a particular consumed dose rather than enumeration of the pathogens that initiated the infection. 'Probability of passage' models cannot be used when the number of microorganisms (or particles) following some process exceeds the number subjected to the process. Accordingly, recovery models in which counting errors may inflate counts relative to the number of particles actually present (*e.g.* Section 4.2.4) are not 'probability of passage' models. 'Probability of passage' models are particularly appropriate for dose-response models because it is not possible for more pathogens to initiate an infection than were originally consumed. Viruses and pathogenic protozoa such as various genotypes of *Cryptosporidium* and *Giardia* can only propagate by infection, so there cannot be growth in their numbers prior to infection. The number of bacterial pathogens could grow by cell division, but such growth could be implicitly included with the host-pathogen interactions: the probability of passage in this case is the probability that each bacterium or any of its progeny survive to initiate infection.

Assumptions of 'probability of passage' models used for log-reductions

It is possible to model the probability of passage for a treatment process, and to relate this probability of passage (p) to log-reduction (ρ) using the relationship $\rho = -\log_{10}(p)$. Accordingly, use of 'probability of passage' models has been extended to removal scenarios associated with drinking water treatment (*e.g.* Teunis *et al.* 1997; 1999; 2009; Teunis and Havelaar, 1999; 2002; Medema *et al.*, 2003). In Teunis *et al.* (1999), it is proposed that a binomial or beta-binomial 'probability of passage' model can be used to model the probability of passage among paired samples (*e.g.* samples in which the particles are enumerated before and after treatment). "Strictly speaking, these two methods only apply to the situation that a total number of organisms and a number of organisms with a certain property of interest have been counted in physically the same samples" (Teunis *et al.*, 1999). This model is convenient because the beta distribution, if it is used in datasets with multiple pairs of counts, describes variability in the probability of passage. This model is acceptable if each sample is wholly and precisely enumerated before and after treatment, all samples are treated identically (and independently), and the beta distribution suitably describes the variability in the probability of passage among samples. These assumptions, however, are impractically restrictive because treatment process experiments generally do not yield paired data of this type and because it may be unrealistic to assume that microscopic particles or microorganisms in a water sample can be enumerated exactly.

A more typical example of pairing (referred to as 'matching' herein to distinguish it from the pure form of pairing described in the preceding paragraph) would be to have an enumerated subsample from some larger volume before and after treatment. The subsample before treatment can then be used to infer the number of particles subjected to the treatment, and the subsample after treatment can be used to infer the number of particles that passed the treatment. Unlike paired counts, in which the particles are wholly and precisely enumerated in the same physical sample before and after treatment, matched counts (which must also be precisely enumerated and consist of subsamples from the same sample before and after treatment) have sub-sampling error. It is plausible for the pre-treatment count to be lower than the post-treatment count due to sub-sampling error (a clear violation of 'probability of passage' models that disregard sub-sampling error). Even this relaxed definition of pairing is generally infeasible because (1) analytical recovery is often not evaluated using sub-samples from the same physical water sample before and after treatment, and (2) it is improbable that any enumeration method for microscopic particles or microorganisms in water yields exactly 100% analytical recovery all of the time (*i.e.* the matched counts are probably imprecise due to analytical error of the enumeration method). Furthermore, counts from a treatment process are neither paired nor matched if subsamples are not from the same finite and well-mixed water volume before and after treatment.

Using a model that assumes pairing or matching will yield probabilities of passage and log-reductions that are excessively variable (*i.e.* more variable than what is true in reality) and possibly non-representative unless all of the following are true:

- counts are obtained from the same finite water volume before and after treatment,
- samples are wholly enumerated before and after treatment, or representative sub-samples are withdrawn from the well-mixed sample before and after treatment and analyzed using a model that addresses sub-sampling error, and
- enumerations are exactly precise, or the pre- and post-treatment enumeration data are analyzed using a model that addresses analytical error (possibly including non-constant analytical recovery).

It is possible to use 'probability of passage' models for fully unpaired samples (*i.e.* by comparing pre-and post-treatment concentrations that are not paired) if the variability in the concentrations, sampling errors, and analytical errors (possibly including non-constant analytical recovery) are addressed in the model. A major limitation of these 'probability of passage' models, however, is that the final concentration can never exceed the initial concentration (because the probability of passage would exceed 100%). In the case of unpaired concentrations, it is quite possible for the final concentration to exceed the initial concentration.

Discussion of 'probability of passage' models for log-reductions

Teunis *et al.* (1997) used binomial and beta-binomial 'probability of passage' models for the removal of spores of sulphite reducing clostridia by a treatment process. Few details are provided about the experiment from which the data were generated and the assumptions of the selected models are not discussed. The data are presented in Teunis and Havelaar (1999): 73 count pairs (with each count based on a 100 mL sample) have initial counts ranging from 2 to 2125 and final counts as high as 18 (though all but 6 of the final counts were non-detects). It is argued (Teunis *et al.*, 1997) that 'probability of passage' models are particularly useful because, unlike log-reductions calculated from raw concentration estimates, the models can accommodate non-detect treated water samples. This is true of any probabilistic model that addresses uncertainty in concentrations or probabilities of passage. It is improbable that the reported counts are actually paired (as the models assume) because this would require specific samples to be wholly and precisely enumerated before and after treatment as described above. Neither sub-sampling error (if the matched sub-samples are actually withdrawn from the same treated and well-mixed water) nor possible analytical error in the enumeration method were addressed. As described above, overlooking these errors and using a paired model will result in excessively variable (and possibly non-representative) probabilities of passage.

Teunis *et al.* (1997) used maximum likelihood estimation to evaluate a constant probability of passage (using the binomial model) and parameters of a beta distribution describing variable probabilities of passage (using the beta-binomial model). The results were $p = 0.0016$, $a = 0.255$, and $b = 4.107$, which were converted to decimal reductions (*i.e.* $\log_{10}$-reductions) using $\rho = -\log_{10}(p)$. On the basis of the beta-binomial 'probability of passage' model, it was asserted that the probability of a $\log_{10}$-reduction greater than 25 is approximately 25% (or that there is a 25% probability that the probability of passage is less than 10^{-25}). Such high log-reductions cannot be empirically substantiated unless samples seeded with upwards of 10^{25} microorganisms routinely yield non-detects after treatment. The highest log-reduction supported by any of the pairs (using the detection limit of one over the initial count for non-detects) is only 3.32. This problem may reflect high-reduction extrapolation because of the numerous non-detect data. Alternatively, these high log-reductions may have resulted from using a model that assumes pairing for counts that are not actually paired (in the strict technical sense). Finally, these high log-reductions may indicate that the beta distribution is inappropriate for this type of 'probability of passage' data. Whenever $a < 1$ in a beta distribution, the probability density will increase as the probability of passage approaches zero (*i.e.* the probability density for probabilities of passage that correspond to improbably high log-reductions is greater than for probabilities of passage that correspond to tangibly moderate log-reductions). Similar results were obtained using these models for spore reduction in Teunis *et al.* (1999). It is imperative that these particular models not be used for unpaired counts in which measurement errors (*e.g.* sub-sampling from a larger treated volume or analytical error in the enumeration process) have been ignored, and the appropriateness of the beta distribution to model variable probabilities of passage should be researched further.

Teunis *et al.* (1999) expanded the modelling considerations for the probability of passage associated with spore and virus enumeration data. The binomial and beta-binomial models described above were used only when the post-treatment count was assumed to be a subset of the pre-treatment count in a single wholly enumerated sample. Alternative models were developed to address the situation in which counts are not obtained from the same sample. The 'paired samples' model addresses sub-sampling error in the pre- and post-treatment counts (these are described as matched counts above) and assumes that the samples represent the same water before and after treatment. The 'unpaired samples' model addresses variability in the pre- and post-treatment concentrations using counts that are not from the exact same water samples (in fact the numbers of pre- and post-treatment counts are not necessarily equal). In these models, it is assumed that sub-sampling error is Poisson-distributed and that concentration varies among sampling events according to a gamma distribution. In both cases, the probability of passage was modelled as a constant or using a beta distribution. In some cases, the beta distributions had $a < 1$ and will likely support extremely high log-reduction values. Comparison of the six alternative models is informative for demonstration purposes, but the results cannot be used to allow the analyst to choose models based upon fit. It is imperative that probabilistic models be chosen so that they are representative of the process by which the data are generated: any non-representative model is incorrect (*e.g.* it may attribute variability to the wrong sources and lead the analyst to dangerous conclusions) regardless of fit.

Teunis and Havelaar (1999) disputed the use of the beta distribution for temporal variability in the probability of passage associated with treatment processes because the spore reduction data suggested a bimodal distribution (*i.e.* a distribution typically consisting of low probabilities of passage that has occasional high values that may indicate treatment plant failure). If the data were inconsistent with the strict definition of pairing upon which the model was based, then this problem may also have arisen from unaddressed random errors. For example, if the concentrations vary over time and the data are not strictly paired, then a high probability of passage could be inferred from a high post-treatment count that is falsely paired with an unrelated, low pre-treatment count. In filtration, unpaired counts will also result from detachment (*e.g.* post-treatment counts that are atypically high because of previously accumulated microorganisms that suddenly break through). Sub-sampling errors or analytical errors in the enumeration method could also lead to anomalous probabilities of passage. It would be prudent to rule out short-comings of the model before attributing anomalies to failure of the treatment system. Teunis and Havelaar (1999) used the 'unpaired samples' model from Teunis *et al.* (1999) for data associated with the removal of *Cryptosporidium* oocysts because it was acknowledged that counts before and after storage in a large reservoir are unpaired.

Teunis *et al.* (2009) presented essentially the same 'probability of passage' models as the 'paired samples' and 'unpaired samples' models presented in Teunis *et al.* (1999) for use in the analysis of virus reduction. The main difference was that Bayesian methods (using Markov Chain Monte Carlo) were used to evaluate the unknown model parameters rather than maximum likelihood estimation. These models, however, still assume that the enumeration method has perfect analytical recovery. Teunis *et al.* (2009) did not address the paired binomial and beta-binomial models that were frequently used in earlier publications.

A fundamental limitation of converting 'probability of passage' models to log-reduction models is that negative log-reductions (*i.e.* probabilities of passage exceeding 100%) are not allowed: the treated concentration is not allowed to exceed the pre-treatment concentration. Negative log-reductions can happen when counts are not strictly paired or when uncertainty in pre-treatment and post-treatment concentrations is considered. An additional limitation is that none of these 'probability of passage' models for log-reductions account for the situation in which multiple pre-treatment counts or post-treatment counts are regarded as replicates. Further research is needed to validate the use of 'probability of passage' models to quantify log-reductions and to generalize the models to address analytical errors and replication.

5.4.2 Concentration Ratio Models

A different approach to evaluate a single log-reduction estimate is to compute it directly from the ratio of the two concentration estimates. Emelko (2001) generated a distribution that represents the uncertainty in a log-reduction estimate by evaluating the uncertainty in the concentration estimates (based upon single or replicate data). Uncertainty in the concentration estimates was evaluated using Bayes' theorem and Gibbs sampling (Section 5.3.2). The distribution of log-reduction estimates was generated by concurrently running Gibbs sampling algorithms for the initial and final concentration and computing log-reduction from the two concentration values generated in each iteration. A similar approach was used in Emelko *et al.* (2010a), for both the beta-Poisson and negative binomial models, except that the order of the generated concentration values in each sequence was randomized to calculate the log-reduction distribution (because this yields better mixed sequences of log-reduction values when either of the Gibbs sampling concentration sequences is poorly mixed). This type of approach for evaluating uncertainty in log-reduction estimates is a basic Monte Carlo that addresses uncertainty (rather than variability) in the two concentrations. The uncertainty in log-reduction is evaluated by computing log-reduction from pairs of concentration values that are randomly drawn from their respective posterior distributions. The equivalent integration approach was also presented in Emelko *et al.* (2010a) using Equation 5.25 (in which C_1 is the initial concentration, C_2 is the final concentration, R is the log-reduction, and ρ is a particular value of log-reduction).

$$P(R > \rho) = \int_0^{\infty} P(C_1 = c) \cdot P\left(C_2 > c \cdot 10^{-\rho}\right) dc \tag{5.25}$$

The resulting distribution for log-reduction was referred to as the posterior log-reduction distribution in Emelko *et al.* (2010a). It is, however, not a posterior distribution because the log-reduction analysis is only partially Bayesian;. Log-reduction is not regarded as an unknown parameter in Bayes' theorem. Furthermore, the distribution representing uncertainty in log-reduction is obtained by regular Monte Carlo (or integration) assuming that the two posterior concentrations are independent. The concentrations, however, are not independent because $C_2 = C_1 \cdot 10^{-R}$. Therefore, the generated distribution is not the posterior log-reduction distribution even though it does describe uncertainty in log-reduction. If the dependence between the two concentrations strongly affects their resulting posteriors, then the distribution generated in this way may be a poor approximation of the posterior log-reduction distribution. The results of an analysis using numerical integration to obtain posterior concentration distributions and using Equation 5.25 to evaluate uncertainty in log-reduction is compared to a fully Bayesian approach in Section 5.4.3.

Teunis *et al.* (2009) described an approach to evaluate variability in log-reduction when variability in the initial and final concentrations is modelled using gamma distributions. They demonstrated that the ratio of the two concentrations is F-distributed if C_1 and C_2 are independently gamma-distributed. The model is incorrect, however, because the distribution of C_2 is dependent upon the distribution of C_1 and the distribution of log-reduction. Furthermore, log-reduction itself is not necessarily independent of the initial concentration (*e.g.* Assavasilavasukul *et al.*, 2008). Finally, the model used to obtain the gamma distributions for the two concentrations did not address analytical error (*i.e.* it was assumed that the enumeration method has a consistent analytical recovery of 100%).

5.4.3 Fully Bayesian log-Reduction Models

As described in Section 5.4.2, the post-treatment concentration is a function of the pre-treatment concentration and the log-reduction ($c_2 = c \cdot 10^{-\rho}$). Herein, the probabilistic models that are used to describe measurement errors in the enumeration data (*i.e.* the beta-Poisson and negative binomial enumeration models) are expanded to include the situation in which two concentrations are being estimated and are related by log-reduction. Each concentration may be estimated using single or replicate data. Upon applying Bayes' theorem to the model with a suitable prior $g(c,\rho)$, the resulting posterior log-reduction distribution quantitatively describes uncertainty in log-reduction given the model, available data, and the prior. The models presented herein consider the uncertainty in only a single log-reduction estimate, and do not address variability in log-reduction over time. Further research is needed to develop such a model.

Discussion of the fully Bayesian 'beta-Poisson log-reduction model' and 'negative binomial log-reduction model', and derivation of the conditional posterior distribution functions, is provided in Appendix B.4. Gibbs sampling is used herein to evaluate the posterior log-reduction distribution. The conditional posterior distribution functions for the numbers of particles collected in each sample (in the beta-Poisson model) and the analytical recovery value for each sample are unchanged (except that the reduced concentration $c \cdot 10^{-\rho}$ is substituted where appropriate). In the beta-Poisson log-reduction model, the conditional posterior distributions for c and ρ are represented by Equations 5.26 and 5.27. In these equations, the subscripts $1k$ and $2k$ represent the k^{th} of r_1 or r_2 replicate enumerations of the initial or final concentration respectively. If $g(c|\rho)$ is the improper semi-infinite uniform prior, then $c(\Sigma V_{1k} + 10^{-\rho} \cdot \Sigma V_{2k}) \sim \text{GAMMA}(\Sigma n_{1k} + \Sigma n_{2k} + 1,1)$ can be used to generate c. If $g(\rho|c)$ is the improper infinite uniform prior (because $-\infty < \rho < \infty$), then $10^{-\rho} \cdot c \cdot \Sigma V_{2k} \sim \text{GAMMA}(\Sigma n_{2k} + 1,1)$ can be used to generate ρ. Gibbs sampling, therefore, is very simple in this scenario.

$$Dfc(c) \propto g(c \mid \rho) \cdot \exp\left(-c\left(\sum_{k=1}^{r_1} V_{1k} + 10^{-\rho} \cdot \sum_{k=1}^{r_2} V_{2k}\right)\right) \cdot c \wedge \left(\sum_{k=1}^{r_1} n_{1k} + \sum_{k=1}^{r_2} n_{2k}\right) \tag{5.26}$$

$$Dfc(\rho) \propto g(\rho \mid c) \cdot \exp\left(-c \cdot 10^{-\rho} \cdot \sum_{k=1}^{r_2} V_{2k}\right) \cdot \left(10^{-\rho}\right) \wedge \left(\sum_{k=1}^{r_2} n_{2k}\right) \tag{5.27}$$

The Gibbs sampling algorithm that was used herein is displayed in Figure 5.9. The initial parameter values that were used were $p_{ik} = a \,/\, (a + b)$ and $c_i = \Sigma x_{ik} \,/\, \Sigma V_{ik} p_{ik}$ unless $\Sigma x_{ik} = 0$, in which case Σx_{ik} is assigned a value of 0.5. From these the initial parameters, c and ρ are c_1 and $\log_{10}(c_1 \,/\, c_2)$ respectively. It is possible to use sample-specific recovery parameters (a_{ik}, b_{ik}) and to incorporate uncertainty in these parameters (as shown in Figure 5.9), but neither is done herein for simplicity.

The Table 5.3 enumeration data were analyzed using this Gibbs sampling process with $a = 287.08$, $b = 94.76$ and 30,000 iterations (following a burn-in of 1000 iterations). These results were compared to the partially Bayesian approach described in Section 5.4.2 and Emelko *et al.* (2010a) using posterior distributions for the initial and final concentration obtained by numerical integration. Figure 5.10 shows the distributions describing uncertainty in log-reduction based on the partially Bayesian approach discussed in Section 5.4.2 and the fully Bayesian approach discussed in this section. Figure 5.10b shows that the two approaches yield essentially the same distribution. In some cases, the fully Bayesian approach is no more informative than the somewhat easier partially Bayesian approach because the information about the initial concentration that is provided by the final concentration enumeration data and the log-reduction is not substantially more informative than treating the two concentrations as independent.

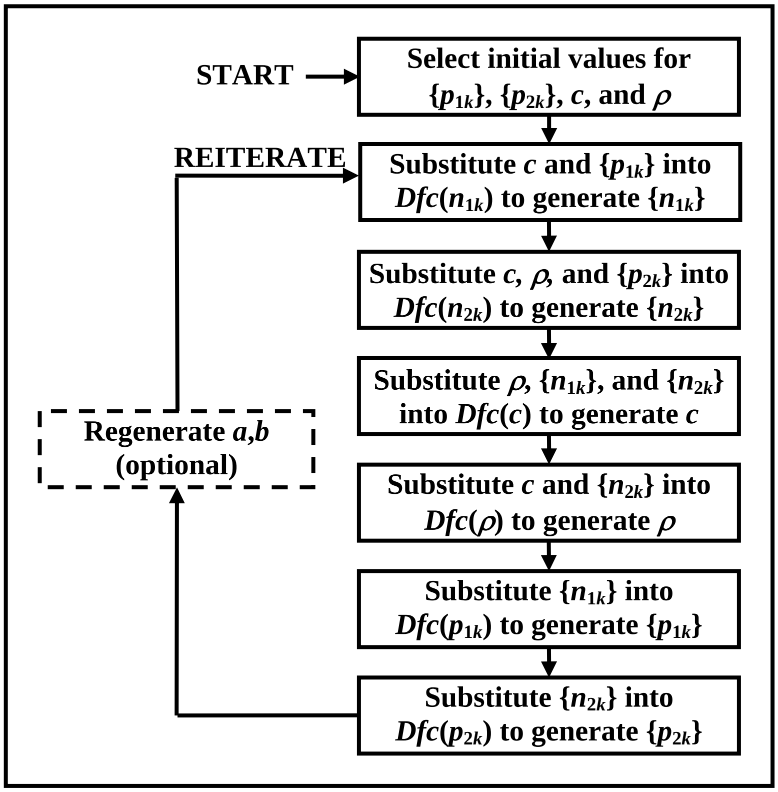

START
Select initial values for $\{p_{1k}\}$, $\{p_{2k}\}$, c, and ρ
REITERATE
Substitute c and $\{p_{1k}\}$ into $Dfc(n_{1k})$ to generate $\{n_{1k}\}$
Substitute c, ρ, and $\{p_{2k}\}$ into $Dfc(n_{2k})$ to generate $\{n_{2k}\}$
Substitute ρ, $\{n_{1k}\}$, and $\{n_{2k}\}$ into $Dfc(c)$ to generate c
Substitute c and $\{n_{2k}\}$ into $Dfc(\rho)$ to generate ρ
Substitute $\{n_{1k}\}$ into $Dfc(p_{1k})$ to generate $\{p_{1k}\}$
Substitute $\{n_{2k}\}$ into $Dfc(p_{2k})$ to generate $\{p_{2k}\}$
Regenerate a,b (optional)

Figure 5.9: Gibbs sampling algorithm for the beta-Poisson log-reduction model

The Gibbs sampling algorithm begins with a burn-in and ends after a user-specified number of iterations. Uncertainty in the parameters of the beta distribution describing non-constant analytical recovery can be addressed by using new estimated parameter values from a posterior distribution (*e.g.* using Gibbs sampling as discussed in Section 4.4.3).

5.4.4 Case Study: Evaluation of Oocyst Removal in Pilot-scale Filtration

Log-reduction is often estimated directly from raw concentration estimates (*i.e.* counts per unit volume) without regard for the measurement errors that make the concentration and log-reduction estimates uncertain (and possibly biased). This uncertainty can be fully addressed (*e.g.* using the method presented in Section 5.4.3) if raw counts and sample volumes are available as well as information about the analytical recovery of the enumeration method. It is plausible that analytical recovery may be different in treated and untreated waters (because the water quality has presumably changed), and so it is advisable to obtain information about analytical recovery in both types of water. If the mean analytical recovery in the two types of water is substantially different, then failing to account for recovery in log-reduction estimates will result in bias.

Huck *et al.* (2001) provided several examples of pilot-scale filtration experiments in which information about analytical recovery was obtained for the treated and untreated water. One small portion of the data that they obtained (as summarized in Table 5.4) is analyzed herein as a case-study. The raw data (*i.e.* counts and sample volumes) associated with these concentration estimates were not provided. Similarly, the seed doses (with an appropriate discussion of seeding methodology) and

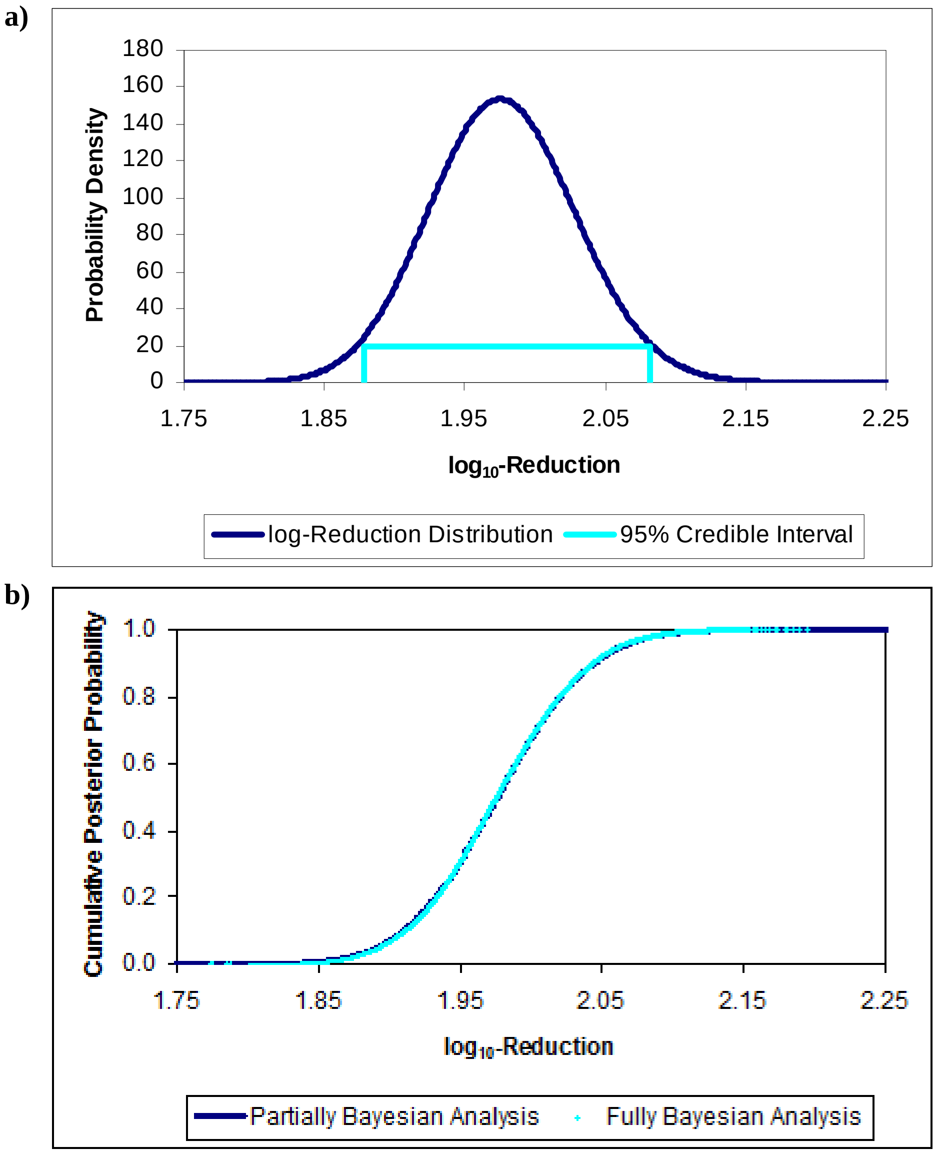


Figure 5.10: Uncertainty in log-reduction in the Table 5.3 data

These figures show (a) the log-reduction distribution generated using independent posterior concentration distributions (generated by numerical integration) and the 95% credible interval (obtained using the fully Bayesian Gibbs sampling approach) and (b) cumulative probability density functions obtained using the partially Bayesian numerical integration approach (that treats initial and final concentrations as statistically independent) and the fully Bayesian Gibbs sampling approach (that treats log-reduction as an unknown parameter in the model by which the two concentrations are related.

Table 5.4: *Cryptosporidium* log-reduction data obtained during stable filter operation*

Sampling Time (min)	Initial Concentration (oocysts/L)	Final Concentration (oocysts/L)	Log-removal (-)
15	9.48×10^4	2.40×10^2	2.6
30	8.36×10^4	1.90×10^2	2.6
40	6.42×10^4	1.90×10^2	2.5
50	5.86×10^4	1.20×10^2	2.7

* Modified from the 7/15/98 stable filter operation experiment data – Table B.10 of Huck *et al.* (2001)

Table 5.5: Unbiased concentration and log-removal estimates

Sampling Time (min)	Initial Concentration (oocysts/L)	Final Concentration (oocysts/L)	Log-removal (-)
15	3.41×10^5	6.50×10^2	2.72
30	3.01×10^5	5.15×10^2	2.77
40	2.31×10^5	5.15×10^2	2.65
50	2.11×10^5	3.25×10^2	2.81

counts were not provided in the recovery datasets. It is not possible to evaluate the accuracy or precision of the reported concentration and log-removal estimates without the raw data from which these estimates were obtained. Given the cost and difficulty associated with obtaining such datasets, it would behove analysts to provide the raw data (where possible) so that uncertainty can be addressed and optimal scientific value can be obtained from the available data. To facilitate the analyses presented herein, hypothetical raw data were inferred from the available information. The hypothetical data used herein are summarized in Appendix D.3.

The recovery estimates in Huck *et al.* (2001) were calculated as a ratio of seeded and 'measured' concentrations. It is assumed that these recovery data are consistent with the beta-Poisson recovery model (Section 4.2.3) because it addresses the situation in which the concentration in seeded samples is known rather than the precise numbers of seeded particles. Estimates of the parameters of the beta distribution describing non-constant analytical recovery were obtained using maximum likelihood estimation (Section 4.4.1). This was conducted separately for the filter influent and filter effluent samples. The filter influent samples were seeded with either 200 oocysts/mL or 10 oocysts/mL, while the filter effluent samples were seeded with approximately 0.2 oocysts/mL. It is assumed herein that the filter influent analytical recovery is independent of the two seeded concentrations. This assumption could be evaluated using the Bayesian equivalent of a hypothesis test as described in Section 4.4.5. The beta distribution parameters obtained by maximum likelihood estimation were $a = 18.78$, $b = 48.75$ ($\mu = 0.2781$, $\sigma = 0.0541$) for the filter influent samples and $a = 24.75$, $b = 42.29$ ($\mu = 0.3692$, $\sigma = 0.0585$) for the filter effluent samples. The analytical recovery in the filter effluent appears to be somewhat higher on average than the recovery in filter influent. The weight of evidence supporting such an assertion could be evaluated using the posterior distribution of the difference between two means (as discussed in Section 4.4.5), but is not evaluated herein.

It is presumed herein that the concentration estimates provided in Table 5.4 were not adjusted for the mean analytical recovery of the enumeration method (*i.e.* they are simply counts per unit volume). Such concentration estimates are biased unless the mean analytical recovery is 100%. Log-reduction estimates, however, will only be biased if they are calculated from concentration estimates that are unequally biased (*i.e.* if the bias in each concentration is equal, then the log-reduction estimate is unbiased). The bias can be calculated as $\log_{10}(\mu_{p2}/\mu_{p1})$, in which μ_{p1} and μ_{p2} are the estimated mean analytical recoveries of the filter influent and filter effluent respectively. In this case, the reported log-removal estimates are biased low by 0.123. The bias is generally much less than 1 because mean analytical recovery is unlikely to vary by an order of magnitude between the two types of water. Table 5.5 summarizes the unbiased concentration and log-removal estimates that reflect the mean analytical recovery of the enumeration method.

To facilitate analysis of the uncertainty in these concentration and log-removal estimates, it was necessary to use hypothetical enumeration data and sample volumes (because these details were not provided in Huck *et al.*, 2001). The hypothetical enumeration data used herein that lead to the reported concentration and log-removal estimates are summarized in Appendix D.3. These data are analyzed herein using the beta-Poisson enumeration model and the fully Bayesian log-reduction analysis approach discussed in Section 5.4.3. The maximum likelihood estimates for the beta distribution parameters that are discussed above are used herein. In each analysis, Gibbs sampling consisted of a burn-in of 1000 iterations followed by an additional 30,000 iterations that were used to represent the posterior distribution. The results are summarized in Figure 5.11.

Figure 5.11a clearly illustrates that concentrations reported as counts per unit volume are biased if mean analytical recovery is not 100%. In this particular example, the low mean analytical recovery causes the concentration estimates to be so biased that they are not even within the associated 95% credible intervals (which represent the most probable actual concentration values supported by the data). In this experiment the biased concentrations are not particularly problematic because the log-removals are of interest rather than the concentrations. In other applications (*e.g.* monitoring and risk analysis for pathogens in drinking water), conventional concentration estimates obtained using methods with low mean analytical recovery will have minimal value because they can be substantially different from the actual concentration that is of scientific and/or regulatory interest. In this particular scenario, the four credible intervals for each set of concentration estimates overlap. It is plausible that the filter influent and filter effluent concentrations were constant and that the decreasing trend in concentration is simply due to random measurement errors. If the two concentrations were actually constant, then it would be possible to evaluate the data as replicates and to obtain a single log-removal estimate.

Figure 5.12b illustrates the bias of the log-removal estimates that did not account for mean analytical recovery. The bias, however, is smaller than the uncertainty in the log-removal estimates (as illustrated by the 95% credible intervals). If the mean analytical recovery is higher in the treated water than in the untreated water, then log-reduction estimates will be biased low. Such an error is

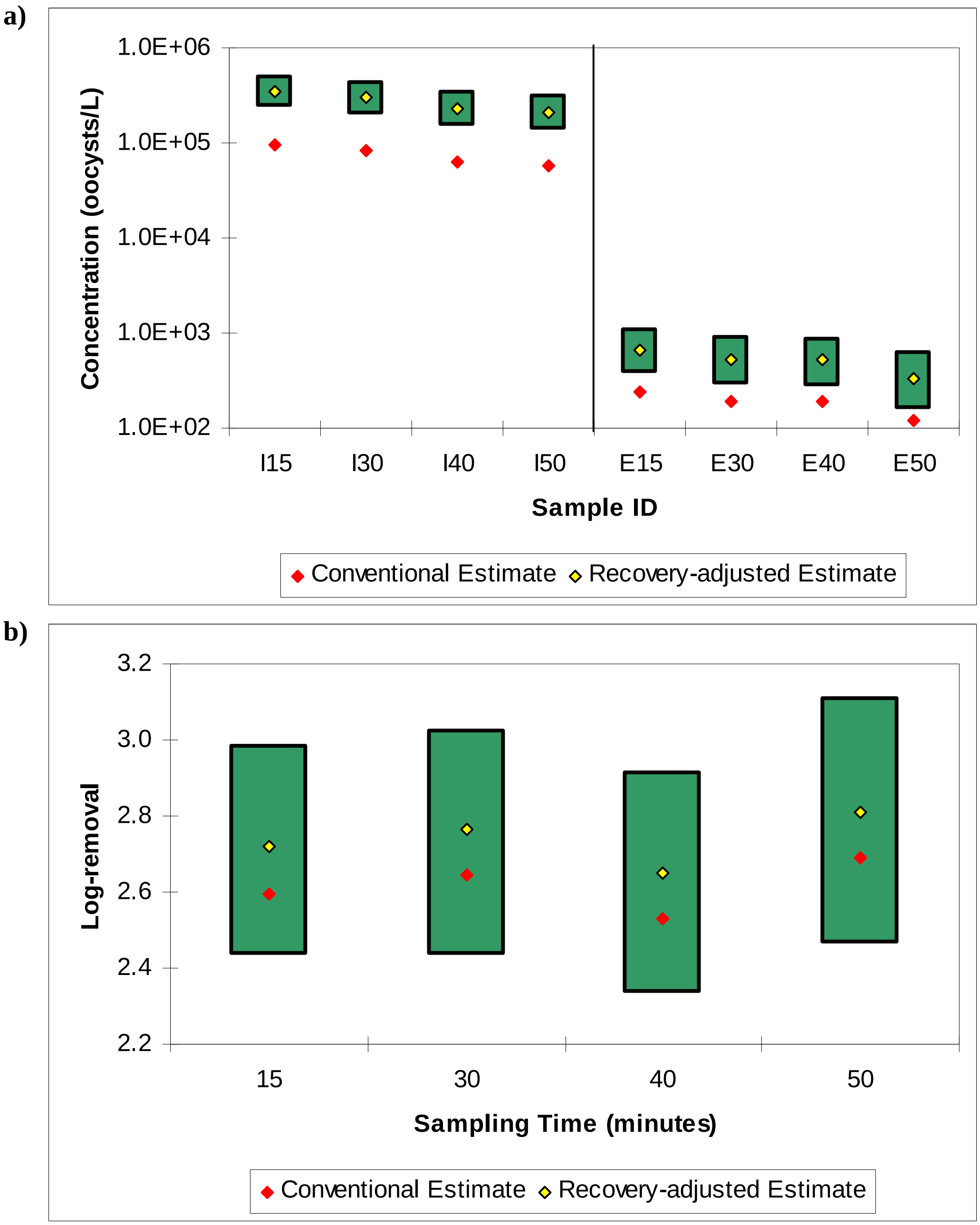


Figure 5.11: Uncertainty in concentration and log-removal estimates
(after Roberson *et al.*, 2010)

These figures illustrate the uncertainty in estimates of (a) the filter influent and filter effluent concentrations and (b) the log-removal. For concentration, the conventional estimate is the count per unit volume and the recovery-adjusted estimate is the count per unit volume divided by the mean analytical recovery of the enumeration method. The conventional and recovery-adjusted estimates for log-removal are calculated using the conventional or recovery-adjusted concentration estimates respectively. The boxes represent 95% credible intervals, which were computed from Gibbs sampling results using the beta-Poisson log-reduction model (Section 5.4.3).

conservative in the case of water treatment because better water will be produced if the log-reduction is higher than what is reported. The credible intervals show that these log-reduction estimates are relatively uncertain. Considering the uncertainty, it can be said (on the basis of a Bayesian analysis and the hypothetical enumeration data in Table D.3) that this stable operation filtration experiment yielded a log-removal that is significantly greater than 2.3 at each of four sampling occasions. It is also plausible that the log-removal was constant given that the four log-removal credible intervals overlap substantially. The slight variations could possibly be due to measurement error rather than variability in log-removal over time.

This dataset was also analyzed in Roberson *et al.* (2010). In those analyses, Gibbs sampling was used and log-reduction was evaluated using the method presented in Section 5.4.2 (rather than the fully Bayesian method used herein). The results were very similar to the results presented in Figure 5.11, which confirms that the method to evaluate uncertainty in log-reduction that was presented in Emelko *et al.* (2010a) and in Section 5.4.2 is often a good approximation of the posterior log-reduction distribution.

5.5 HYPOTHESIS TESTING USING CONCENTRATION ESTIMATES

It is often desirable to use classical statistical tools such as confidence intervals and hypothesis tests to analyze concentration estimates. There are, however, many assumptions and limitations of these tools that often preclude their proper use in this type of application. Section 5.5.1 addresses the use of these conventional tools with enumeration-based concentration estimates. Section 5.5.2 describes how posterior distributions obtained using the Bayesian methods presented herein can be used to facilitate decision-making.

5.5.1 Using Conventional Statistical Tests with Enumeration Data

Conventional parametric statistical tools such as t-tests, chi-squared tests, and F-tests (and their associated confidence intervals) may sometimes be chosen to analyze enumeration-based concentration data. These have several assumptions that often preclude their proper use in the analysis of this type of data. In particular, these tools assume that the data are independent, normally distributed, and have constant variance (*i.e.* it is assumed that each concentration or log-reduction estimate is as precise as the others). Another problem is that these tools address uncertainty in the population mean or population variance of the concentration estimates rather than the actual concentration.

A t-test (Section 3.2.1) addresses uncertainty in the population mean of a set of data. Two possible scenarios may be considered in the case of concentration estimates: (1) fully replicated concentration estimates have been obtained and their population mean is assumed to be the actual concentration, or (2) non-replicate concentration estimates have been obtained and their population mean is assumed to be the actual average concentration (*i.e.* the mean concentration over the space or time that the samples are assumed to represent). The former scenario is addressed in this chapter (*i.e.* uncertainty in a single concentration value given imprecise single or replicate enumeration data) and the latter is addressed for temporally variable concentrations in Chapter 6. The t-test assumes that each of the data (*i.e.* concentration estimates) is independent and identically normally distributed. In the case of estimating the actual concentration from imprecise concentration estimates, this assumes that each estimate has equal measurement error. Accordingly, the t-test cannot be used if samples have different volumes or different non-constant analytical recovery distributions because some estimates would be more precise than others. Furthermore, these estimates must be approximately normally distributed: therefore, the counts modelled in Sections 5.1.1 or 5.1.2 must also be approximately normally distributed. It was demonstrated in Emelko *et al.* (2010a) that beta-Poisson and negative binomial model count distributions are not approximately normally distributed when counts near zero are commonly obtained (*e.g.* Figure 5.12). The assumptions are more restrictive when concentration is a variable rather than a constant because it may be difficult to assert that the estimates are independent, normality of the estimates will depend upon the distribution of concentration itself as well as the measurement errors, and the data are unlikely to have constant variance (because the precision of concentration estimates is a function of concentration, as demonstrated in Section 8.2.3). In concept, data that are incompatible with the assumptions of normality and constant variance can be transformed to satisfactorily meet these assumptions; however, datasets are often insufficiently large to choose and validate an appropriate transformation. Finally, these tests will not work if only one datum is available or if the sample standard deviation of the concentration estimates is zero (*e.g.* if all estimates have the same count).

Similarly, a test upon the variance can evaluate the variability among imprecise replicate concentration estimates or the variability among non-replicate concentration estimates. These tests are subject to the same issues as the t-tests described above. Tests upon the variance, however, are more sensitive to departures from normality than tests upon the mean (Box, 1953). A simple diagnostic for severe departures from normality is the P-value (α in Equation 5.28) associated with an alternative hypothesis that the concentration is greater than zero. “The normal approximation is obviously not valid if the probability of a negative concentration is not very small” (Emelko *et al.*, 2010a). In addition to the assumptions described above, these tests will not work if only one datum is available or if the sample standard deviation of the concentration estimates is zero (*e.g.* if all enumerations yield the same count).

$$\alpha = P\left(T < \frac{0 - \bar{c}}{s_c \big/ \sqrt{n}} \right) \tag{5.28}$$

The most important issue surrounding the statistical analysis of concentration estimates is bias. If the mean analytical recovery of an enumeration method is not 100%, then concentration estimates calculated using the count per unit volume are biased (*e.g.* as illustrated in Figure 5.11). The population mean of biased concentration estimates is scientifically irrelevant and may be quite different from the actual concentration that is of interest. Accordingly, any statistical analysis upon concentration estimates must account for analytical recovery. “The enumeration data must be

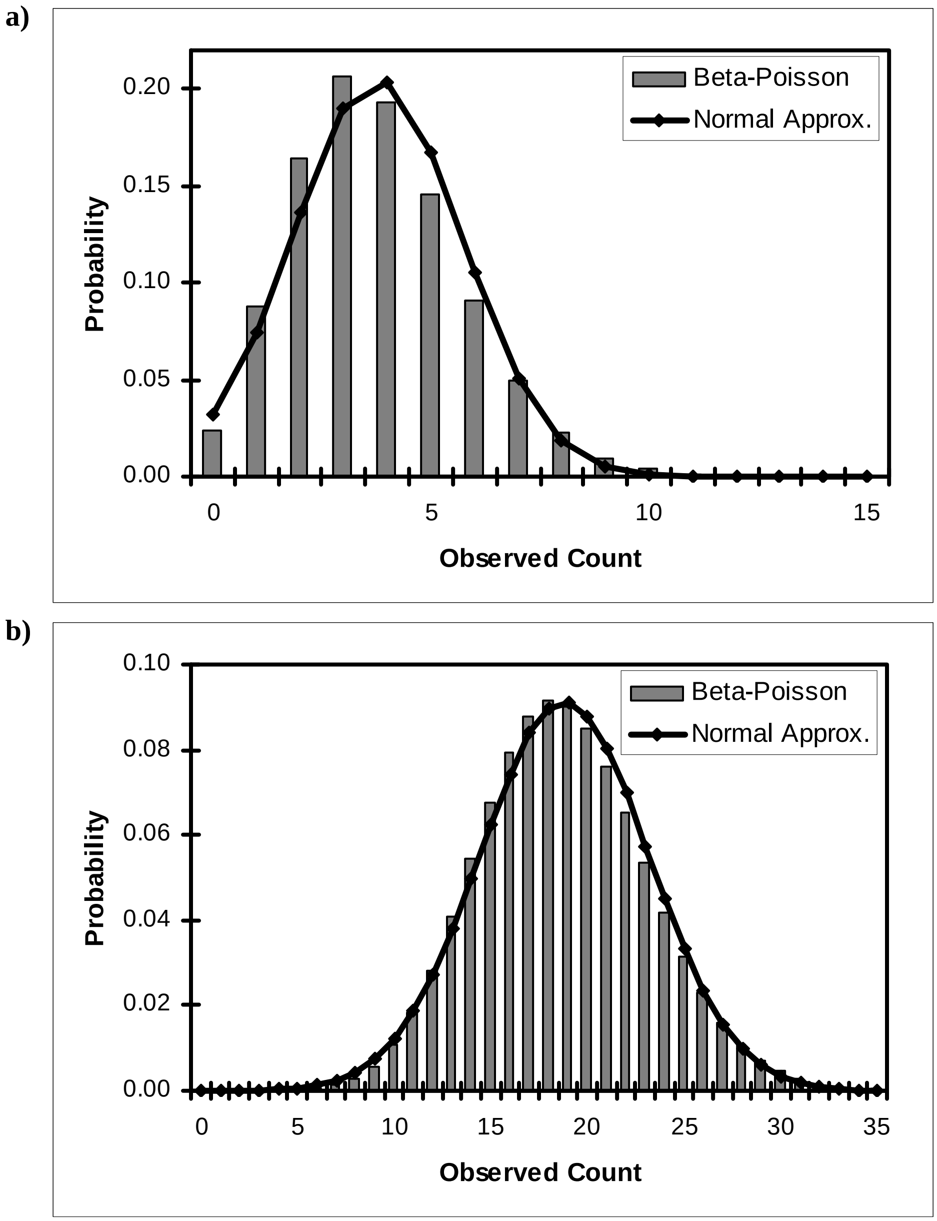


Figure 5.12: Normality of replicate beta-Poisson enumeration data
(Emelko *et al.*, 2010a)

These figures illustrate the normality of count data that follow the beta-Poisson enumeration model (Equation 5.3). Enumeration data are not normally distributed when the counts are often near zero, but the approximation gets better as counts get higher. These examples use beta distribution parameters a = 287.08, b = 94.76, concentration c = 0.5 (particles per unit volume), and sample volumes of (a) V = 10 and (b) V = 50.

converted to concentration estimates and corrected for mean analytical recovery using the equation $\hat{c} = x / V\mu_p$, where x is the observed count, V is the sample volume, and μ_p is the mean analytical recovery" (Emelko *et al.*, 2010a). Concentration estimates can also be calibrated using internal seed recovery estimates. The variability among replicate concentration estimates is the measurement error; however, the variability among non-replicate concentration estimates (*i.e.* when concentration is assumed to be variable) represents both measurement error and variability in concentration. Accordingly, any test upon the variance of non-replicate concentration estimates is biased high because of unaddressed measurement errors. This also reduces the power of hypothesis tests upon the mean to reject false null hypotheses. The same issues were addressed with respect to the variability of recovery estimates in Section 4.4.5.

Many of the issues addressed in this section also apply to log-reduction estimates. Log-reduction estimates are inherently uncertain and can also be biased as demonstrated in Section 5.4.4. When conducting a test upon log-reduction estimates, there are many issues that must be considered: for example, whether or not (1) the estimates are unbiased, (2) the estimates are normally distributed, (3) the estimates are equally precise, (4) the estimates are independent, and (5) the variability among estimates includes variability in log-reduction itself (in addition to ever-present measurement error).

5.5.2 Using Posterior Distributions as a Decision-making Tool

The posterior distributions for concentration and log-removal estimates that are discussed in Sections 5.3 and 5.4 respectively can be used as a robust decision-making tool. These are not subject to the numerous limitations of classical confidence intervals and hypothesis tests. The posterior concentration distribution, for example, provides a complete description of what the analyst is entitled to believe about the actual value of concentration given the available data, the model, and the prior. It is not necessary for replicate counts to be normally distributed or for the data to have constant variance, and uncertainty can even be quantified without replication. Furthermore, the estimated concentrations will not be biased because the model accounts for imperfect analytical recovery. Finally, the Bayesian approach provides a framework for incorporation of prior information (representing results from previous experiments or the analyst's subjective beliefs).

The use of posterior distributions to facilitate decisions is relatively straightforward. The weight of evidence supporting a particular alternative hypothesis can be read directly from the posterior distribution. For example, the posterior distribution for a single concentration or log-reduction estimate can be used to report the probability that the concentration or log-reduction is greater than some regulatory threshold. The posterior distribution of the difference between two independent concentrations can be evaluated by integration or Monte Carlo from the two posterior concentration distributions (as was done to evaluate log-reduction in Section 5.4.2). If concentration varies (which is addressed in Chapter 6), then a posterior distribution can be generated for the mean or variance of concentration. When comparing two concentration distributions, the posterior of the difference between two concentrations or the ratio of two variances can be evaluated using the same methodology as was presented for analytical recovery in Section 4.4.5.

The uncertainty in an estimated value with an associated posterior distribution can be graphically illustrated using credible intervals (*e.g.* Figure 5.11). Credible intervals provide a much more informative display of uncertainty than error bars that are based upon the standard deviation, the standard error of the mean, or percentiles. The standard deviation tells nothing about the skewness of the data and can be biased high due to measurement errors (when illustrating uncertainty in the mean of non-replicate concentration estimates). Furthermore, error bars based upon the standard deviation cannot be provided unless every measurement (*e.g.* concentration estimate) is fully replicated with equal precision. The same issues apply to the standard error of the mean. The standard error, however, can be even less informative because it does not illustrate the standard deviation of fully replicated estimates (unless the number of data upon which it is based is reported). Error bars based upon percentiles are useful for large datasets because they illustrate the spread and skewness of the data. All of these common types of error bars illustrate the scatter of the data, but they do not illustrate uncertainty in a mean, standard deviation, difference between two means, *etc.*

If the multiple issues discussed in Section 5.5.1 have been appropriately addressed, then classical confidence intervals (*e.g.* on the mean, the variance, the difference between two means, *etc.*) can provide a useful graphical display of uncertainty. Confidence intervals, however, "depend only on the data and do not incorporate knowledge of the method by which the data were collected (*e.g.* knowledge that random sampling error follows a Poisson distribution or about the variability in analytical recovery observed in a recovery experiment)" (Emelko *et al.*, 2010a) and cannot reflect prior information like credible intervals. Credible intervals describe what the parameter might be (with specified probability) given the model, prior, and available data, while statements of probability about an unknown parameter cannot be made from confidence intervals.

Posterior distributions are particularly useful because they provide a complete description of uncertainty in the parameter. In contrast, point estimates provide no information about uncertainty and interval estimates only provide a range of plausible values. Moreover, posterior distributions provide a convenient approach to incorporate uncertainty about a particular parameter into other probabilistic models.

5.6 FACTORS THAT AFFECT UNCERTAINTY IN CONCENTRATION AND LOG-REDUCTION ESTIMATES

Knowledge about the measurement errors in the enumeration of microorganisms and discrete particles can be used to facilitate more appropriate statistical analyses of data that have already been collected (*e.g.* as discussed in Sections 5.3-5.5) or to design experiments that generate more informative data. In Emelko *et al.* (2008), credible intervals for various hypothetical scenarios were compared to investigate the impacts of various factors upon uncertainty in concentration and log-reduction estimates. The factors that were addressed in the analysis of concentration estimates were the obtained particle or microorganism count (Section 5.6.1), the sample volume (Section 5.6.2), replication (Section 5.6.3) and the non-constant analytical recovery distribution (which is addressed in Sections 5.6.1 and 5.6.2). The effect of the obtained counts and non-constant analytical recovery distribution upon uncertainty in log-reduction estimates (Section 5.6.4) was also addressed. A more efficient approach to evaluate the impacts of various factors upon the uncertainty of concentration estimates (using variance decomposition) is presented in Chapter 8.

5.6.1 Effect of the obtained count and analytical recovery upon concentration estimates

In general, higher counts yield more precise concentration estimates. Samples with higher counts yield more precise concentration estimates "because a difference of one or two observations (*e.g.* organisms lost during enumeration) has less effect on the estimated concentration than in samples with relatively low counts" (Emelko *et al.*, 2008). 95% credible intervals were calculated using the numerical integration approach (Section 5.3.1) and the beta-Poisson enumeration model for a selection of count values between 1 and 1,000 and a sample volume of 100 L. Two sets of beta distribution parameters were used: (1) $a = 6.739$, $b = 19.9$, $\mu = 0.253$, $\sigma = 0.083$, and (2) $a = 92.43$, $b = 33.79$, $\mu = 0.732$, $\sigma = 0.039$. The results are plotted in Figure 5.13.

A credible interval is a representation of the uncertainty in an estimated parameter such as concentration. In many cases, it is the width of the credible interval relative to the parameter value that is of interest. For example, an error of 1 oocyst/L associated with an estimated concentration of 10 oocysts/L may be important while the same error associated with an estimated concentration of 1,000 oocysts/L would be regarded as trivially small. Accordingly, the computed credible intervals for concentration are plotted in log-scale in Figure 5.13. The width of the credible intervals in log-scale can be regarded as a relative error (which decreases as counts increase) rather than an absolute error (which increases as counts increase). This result is related to random sampling error because samples with low counts yield more uncertain representations of the source as a whole. In the Poisson model for random sampling error, the absolute variance of counts (λ = concentration × volume) increases as the concentration or volume increase, but the relative standard deviation $\lambda^{-0.5}$ decreases. Some enumeration methods (particularly plating methods) recommend count thresholds (below which counts are considered to be too low to yield reliable concentration estimates). A threshold count of 10 might be suggested based upon the results shown in this figure.

Concentrations are often estimated from enumeration data using the count per unit volume. The counts per unit volume are plotted in Figure 5.13 to illustrate that they are biased concentration estimates. When the counts are low, the bias is small and the credible intervals contain this concentration estimate; however, these concentration estimates fall outside the credible intervals as the counts get higher. The first series of credible intervals represents an enumeration method with low mean analytical recovery (approximately 25%), so the bias is much greater than what is observed in the second series (which has a mean recovery of approximately 73%). Further discussion of the bias in concentration estimates that are not adjusted for analytical recovery is provided in Section 8.2.1.

The figure also suggests that the relative error in concentration estimates has a minimum (*i.e.* that further increases in the count will not reduce the relative error). This minimum appears to be related to the variability of analytical recovery because the first series (with a standard deviation of 8.3%) has

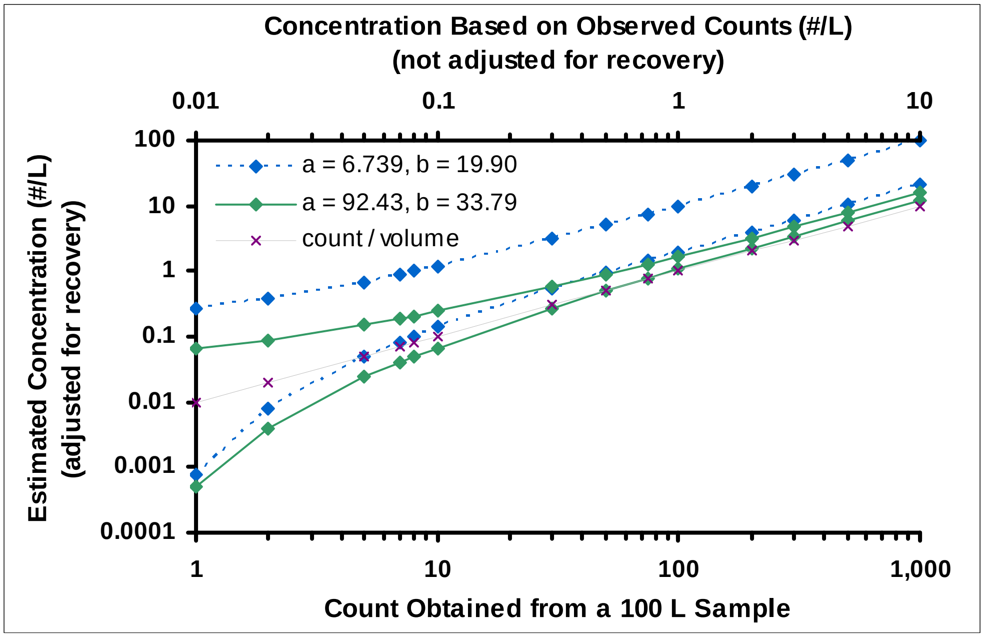


Figure 5.13: Effect of obtained count and analytical recovery upon concentration credible intervals (modified from Emelko *et al.*, 2008)

This figure illustrates 95% credible intervals for concentration with various hypothetical counts of analyte particles in a 100 L sample. The y-axis represents the values of concentration that are supported by the data considering the measurement errors in the enumeration method. The secondary x-axis shows the biased concentration estimates that are obtained when counts are divided by their respective sample volumes without consideration of analytical recovery.

wider credible intervals than the second series (which has a standard deviation of 3.9%). The first series also appears to converge upon its minimum relative error more rapidly (*i.e.* at a lower count). In Emelko *et al.* (2008), it is suggested that the relative error in the second series tapers more rapidly and to a lower minimum value than the first series because more of the measurement error is attributable to sampling error. The random sampling error in a concentration estimate becomes trivially small as the count increases; however, non-constant analytical recovery ensures a residual measurement error as the random sampling error becomes negligible.

It is argued in Emelko *et al.* (2008) that "sample concentrations and volumes resulting in high counts are needed to realize substantial benefits from improvements in [analytical] recovery". Research to improve the analytical recovery of enumeration methods will have little or no impact upon the uncertainty in concentration estimates unless obtained counts are usually sufficiently large that the variability in analytical recovery is the leading cause of error. Improving mean analytical recovery is helpful because higher recovery will lead to higher counts, which has been demonstrated herein to yield more precise concentration estimates, and because analysts who ignore analytical recovery will report less biased concentrations. A method in which analytical recovery is higher on average, but also more variable, can potentially result in more uncertain concentration estimates. These concepts are mathematically proven using variance decomposition in Section 8.2.2. In some cases, simple experimental design enhancements along with the proper use of statistics may go much farther towards improving results than developing methods with better analytical recovery.

5.6.2 Effect of the sample volume and analytical recovery upon concentration estimates

The preceding section addressed the effect of sample counts upon credible intervals for concentration estimates. The count obtained from a sample, however, is a random variable that can only be controlled by changing the concentration, sample volume, or analytical recovery of the enumeration method. This section specifically addresses the effect of increasing the sample volume (while the count per unit volume is held constant). 95% credible intervals were calculated using the numerical integration approach (Section 5.3.1) and the beta-Poisson enumeration model for a selection of count values between 1 and 1,000 (with corresponding sample volumes between 10 and 10,000 L). The same two beta distributions were used to evaluate the effects of analytical recovery. The results are plotted in Figure 5.14.

These results are similar to the preceding analysis. If analytical recovery is independent of the sample volume, then higher volumes will lead to higher counts (on average) and less uncertain concentration estimates. This is because increasing the sample volume will decrease the random sampling error. The count per unit volume (0.1 particles/L) is plotted in Figure 5.14 and is clearly

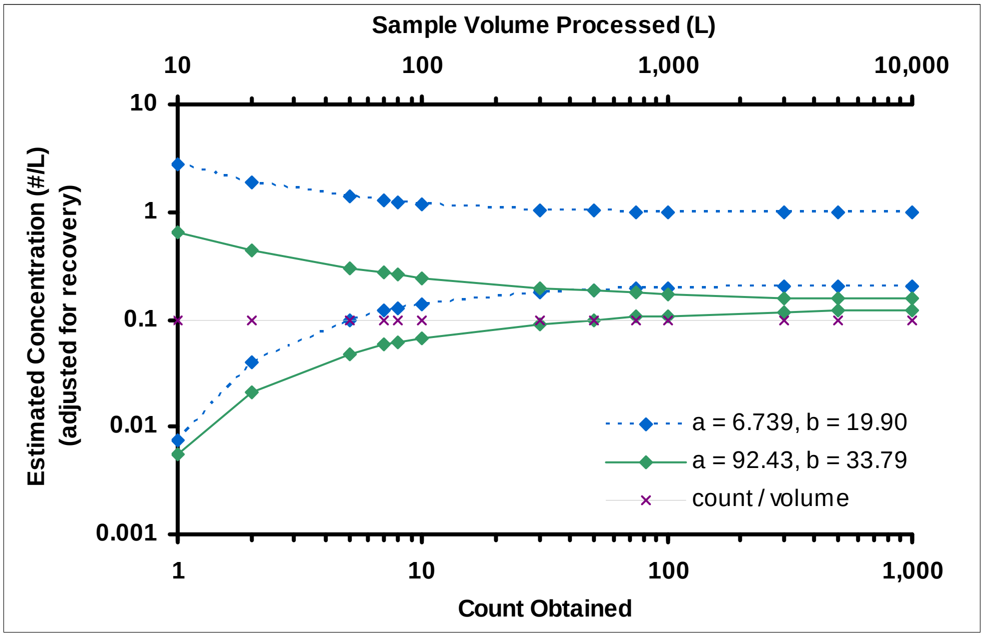


Figure 5.14: Effect of sample volume and analytical recovery upon concentration credible intervals (modified from Emelko *et al.*, 2008)

This figure illustrates 95% credible intervals for concentration with various processed sample volumes in which a hypothetical count of 1 particle for every 10 L is obtained (*i.e.* the conventional concentration estimate is consistently 0.1/L). The y-axis represents the values of concentration that are supported by the data considering the measurement errors in the enumeration method. The secondary x-axis shows the sample volume from which each count was obtained.

biased low, particularly for the method with lower mean analytical recovery. Once again, the relative error tapers faster and to a lower minimum in the series that has less variable analytical recovery. These results demonstrate that more precise concentration estimates can be obtained by increasing sample volumes (where feasible and when random sampling error is a major contributor to overall error). The same results would be true if the concentration were varied with fixed volume and the selected counts were proportional to the concentration; however, the concentration is generally not within the analyst's control except in seeding experiments. The effect of sample volume and concentration upon uncertainty in concentration estimates is also addressed in Section 8.2.1.

5.6.3 Effect of replication upon concentration estimates

When a measurement process (such as estimating concentrations from enumeration data) yields imprecise data, the mean of several data is a more precise estimate of the true value than the individual data. This does not, however, overcome inaccuracy if the data are biased (*e.g.* if concentration estimates do not account for analytical recovery). Emelko *et al.* (2008) evaluated the effect of replication upon 95% credible intervals for concentration by comparing scenarios with 1, 2, 4, 8, or 16 equal-volume replicates that had a consistent total volume of 100 L. In each case, count data were simulated using the beta-Poisson enumeration model (Equation 5.1). Three simulations were completed for each number of replicates because the credible interval depends upon the counts and the counts are random. Figure 5.15 shows the results that were obtained using the recovery parameters a = 6.739, b = 19.90 and two different concentrations.

Figure 5.15a shows that dividing the total volume into replicates yields less uncertain concentration estimates when the concentration is 10 particles/L, while Figure 5.15b does not show any reduction in uncertainty associated with replication when the concentration is 0.05 particles/L. Similar analyses using the recovery parameters a = 92.43, b = 33.79 (not shown) indicated a much smaller effect of replication. The reason is that the beta-Poisson distribution of counts becomes approximately Poisson-distributed when the variability in analytical recovery is small or the random sampling error is large (*e.g.* when the concentration is small). In the case of replicate counts with equal volumes, over-dispersion with respect to the Poisson distribution can arise from non-constant analytical recovery (as discussed in Section 5.2.1). The over-dispersion can be minimal, however, when random sampling and analytical error (which are combined into a single Poisson distribution in Equations 5.2 and 5.4) is the leading contributor to variability in the counts (as opposed to non-constant analytical recovery). Haas (1993) demonstrated that replication of Poisson-distributed counts is equivalent to taking a single larger sample because the counts are additive (*e.g.* replicate counts are collectively as informative about concentration as a single sample with the same total volume). Accordingly, replication (as opposed to taking a single larger sample) is only worthwhile when non-constant analytical recovery causes counts to be substantially over-dispersed (or when it is infeasible to process such a large volume in a single sample). Proof of this relationship between non-constant analytical recovery and the effect of replication upon uncertainty in concentration estimates (using variance decomposition) is given in Section 8.2.2.

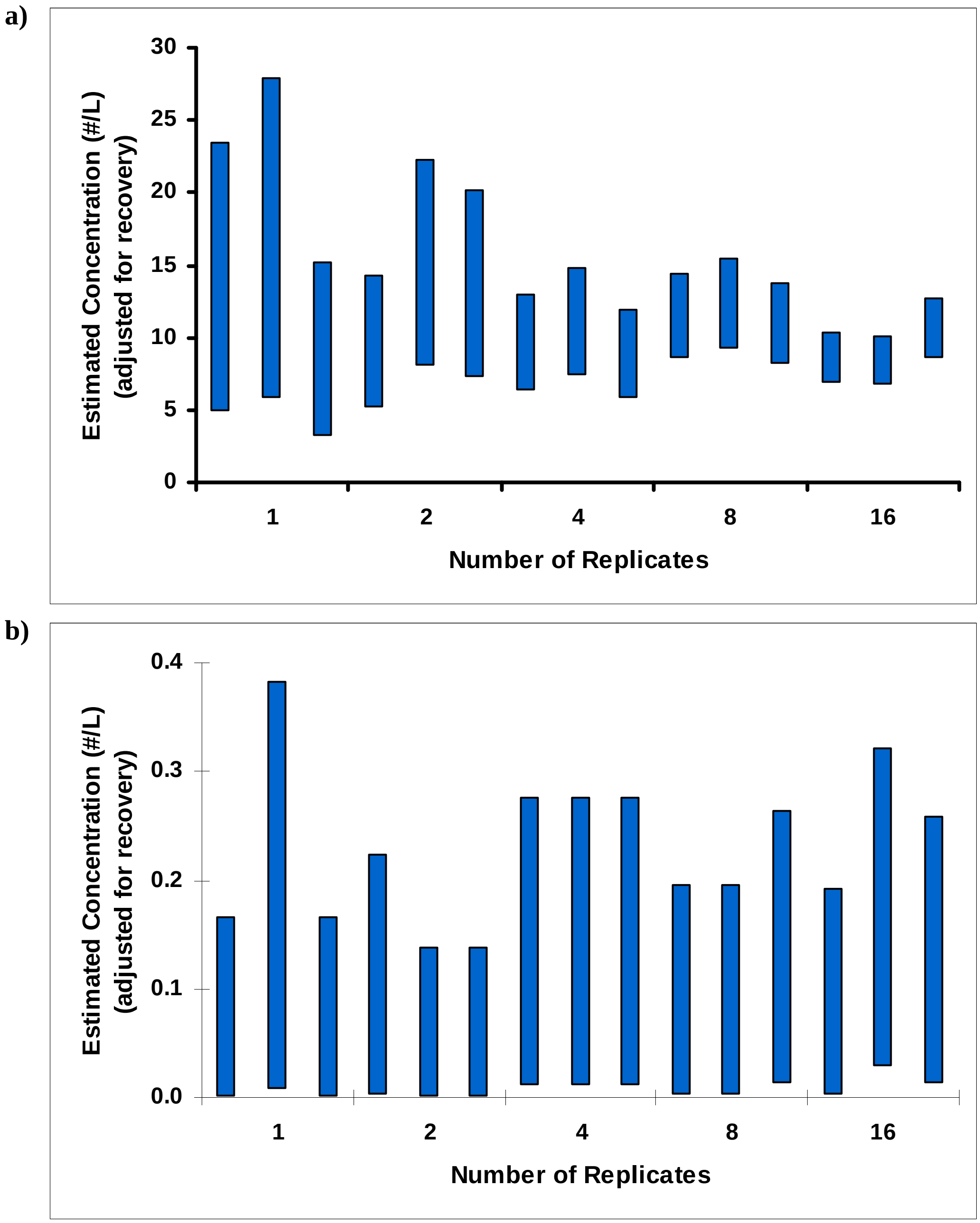


Figure 5.15: Effect of replication upon concentration credible intervals
(modified from Emelko *et al.*, 2008)

These figures illustrate 95% credible intervals for concentrations estimated from various numbers of replicates with a combined volume of 100 L. For each number of replicates, three sets of counts were simulated in accordance with the beta-Poisson enumeration model (Equation 5.1) using a = 6.739, b = 19.90 and a concentration of (a) c = 10 particles/L and (b) c = 0.05 particles/L.

5.6.4 Effect of the obtained count and analytical recovery upon log-reduction estimates

Log-reduction is the ratio of the logarithm of two concentrations. Accordingly, imprecise concentration estimates lead to imprecise log-reduction estimates. A detailed investigation of the factors that affect uncertainty in log-reduction estimates is beyond the scope of this thesis; in general, obtaining more precise concentration estimates will reduce the uncertainty in log-reduction estimates. Emelko *et al.* (2008) evaluated uncertainty in log-reduction estimates when (1) a single 1 L influent sample is collected, (2) a single 100 L effluent sample is collected, (3) the two resulting counts were the same, and (4) the filter influent and filter effluent have the same non-constant analytical recovery distribution. The results of statistical analyses using various counts from 1 to 1,000 (and using the two non-constant analytical recovery distributions that were used in the preceding analyses) are plotted in Figure 5.16. The uncertainty in the log-reduction estimates was evaluated using the method described in Section 5.5.2 (with numerical integration) to yield approximate 95% credible intervals. The fully Bayesian approach that was discussed in Section 5.5.3 would yield the actual 95% credible intervals.

It was found in Section 5.6.1 that higher counts yield more precise concentration estimates; accordingly, Figure 5.16 shows that higher counts yield less uncertain log-reduction estimates. The counts can be increased, thereby improving the precision of the log-reduction estimates, by increasing the concentrations, increasing the sample volumes, or increasing the mean analytical recovery. In

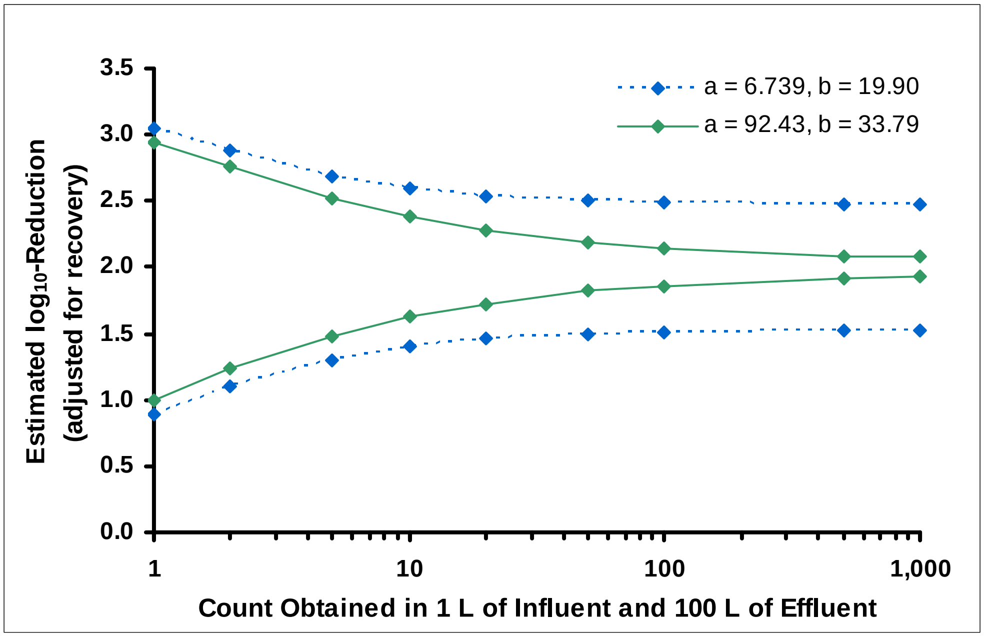


Figure 5.16: Effect of obtained count and analytical recovery upon uncertainty in log-reduction (modified from Emelko *et al.*, 2008)

This figure illustrates 95% probability intervals for log-reduction with various hypothetical counts of analyte particles which are obtained in 1 L of influent and 100 L of effluent (*i.e.* the same count is obtained in one sample from each source). The y-axis represents the values of log-reduction that are supported by the data considering the measurement errors in the enumeration method (using the approach presented in Section 5.4.2).

general, experiments that are conducted to evaluate log-reduction should use seeded influent concentrations and sample volumes that will lead to relatively high counts in the influent and the effluent. It is possible, however, that the seeded concentration may affect the log-reduction of a treatment process (*e.g.* Assavasilavasukul *et al.*, 2008). As described in Section 5.5.4, log-reduction estimates that ignore analytical recovery are only biased if the influent and effluent have different mean analytical recoveries. Therefore, the point estimate of log-reduction associated with each hypothetical scenario shown in Figure 5.16 would be 2.0 (whether the concentration estimates are biased or adjusted for mean analytical recovery). The relative error of the log-reduction estimates tapers faster and to a lower minimum value for the second series (as per Figures 5.13 and 5.14). Once again, this is because the first series has more variable analytical recovery.

5.7 PRACTICAL IMPLICATIONS

The count per unit volume is an uncalibrated (*i.e.* potentially biased) estimate of concentration. The analytical recovery of the enumeration method is the means by which concentration estimates are calibrated to actual concentrations. Unbiased concentration estimates can be obtained by dividing the count per unit volume by the mean analytical recovery of the enumeration method or by a sample-specific recovery estimate. This is demonstrated by Monte Carlo in Section 6.2 and proven by mathematical expectation in Section 8.2.1. In this chapter calibration is only illustrated in Figure 5.11. Calibrating the count per unit volume to an actual concentration using information about the analytical recovery of the enumeration method resolves bias in concentration estimates, but it does not resolve the imprecision. Unbiased concentration estimates are still uncertain because of inescapable random measurement errors in the enumeration process.

Quantitative analyses of single or replicate enumeration data to estimate the concentration of microorganisms or discrete particles in the source must address measurement error. Accordingly, probabilistic models were developed herein to describe the count obtained as a random variable that depends upon the concentration, sample volume, and non-constant analytical recovery distribution. These models were used to develop quantitative analysis tools that address measurement errors in concentration estimates. In particular, Bayes' theorem was implemented to provide a quantitative description of uncertainty in concentration estimates given a model that describes the measurement errors in the enumeration data. The presented Gibbs sampling algorithm is a robust and relatively easy approach to evaluate posterior distributions that describe uncertainty in concentration estimates. The principal limitation of this quantitative analysis approach is the need for suitably representative information about the variability of the enumeration method. This issue is not specific to this method, however, because any appropriate statistical analysis of concentration estimates depends upon information about analytical recovery. The only situation in which it is acceptable to ignore analytical recovery in enumeration data analysis is if the analytical recovery is consistently very close to 100%; in this situation, proper statistical analysis would only have to address random sampling error. Similar issues apply to the estimation of log-reduction because log-reduction is estimated from the ratio of two concentration estimates.

Given the probabilistic modelling framework that is used throughout this thesis, there are three ways in which the effects of various factors upon the precision of parameter estimates can be evaluated: Monte Carlo simulation, variance decomposition, and comparison of credible intervals. The approach used in this chapter was comparison of credible intervals that are obtained from various hypothetical scenarios and parameter values. Compared to Monte Carlo simulation (such as the approaches used in Sections 4.3 and 6.2, this approach is very inefficient and is not recommended. Variance decomposition (Chapter 8) is an even easier approach. Nonetheless, the analyses conducted in Section 5.6 demonstrate that increasing the sample volume (where possible) will often enable collection of more precise concentration estimates. Increasing the sample volume is least impactful in methods with highly variable analytical recovery. It was also found (in general) that it is helpful to choose sample volumes that will yield counts greater than 10 and that replication (*i.e.* enumerating several smaller samples rather than one large sample) is most beneficial when analytical recovery is highly variable.

Conventional statistical confidence intervals and hypothesis tests are often inappropriate for analysis of concentration data because the data are often not normally distributed, may not have constant variance, and are often too few in number to enable transformation. T-tests (and analogous confidence intervals) based upon the mean of a set of replicate unbiased concentration estimates are potentially acceptable if the normality assumption is met because the estimated mean would be the actual concentration and the standard deviation would address measurement error. If concentration varies randomly among a set of concentration estimates, however, then conventional hypothesis tests and confidence intervals should not be used. This is because (1) conventional analyses of the variance would be biased due to unaddressed measurement error in the concentration estimates and (2) the over-estimated variability will reduce the power of hypothesis tests to classify a difference in the mean concentration from the null hypothesis as statistically significant.

An additional conclusion from this research is that it is inappropriate to assert that microorganisms are clustered in a source based solely upon counts that are over-dispersed with respect to the Poisson distribution. This is especially true in the case of non-replicates (*i.e.* samples in which the concentration or enumerated volume may vary), but is also true for replicates because variability in analytical recovery can cause data to be over-dispersed. Finally, it was demonstrated that reported concentration estimates that are not associated with counts and sample volumes (or an appropriate measure of uncertainty in the recovery estimate) have reduced scientific value because there is no justification of the legitimacy and reproducibility of the results and conclusions. In such cases, the precision of the reported values cannot be assessed and it is not possible to subsequently conduct a proper statistical analysis of the data. Moreover, information about the analytical recovery of the enumeration method should be reported along with the data.

Chapter 6

Quantitative Analysis of Temporally Distributed Concentration Data: Enhancing Regulatory Decisions and Risk Analysis

"Variability among concentration estimates can arise from spatial or temporal variability in concentration or from measurement errors (*i.e.* differences between concentration estimates and the actual concentrations that they represent). Consequently, not all of the variation among concentration estimates can be attributed to variability in concentration: some of this variability is due to measurement errors" (Schmidt and Emelko, 2010a). Any model that addresses variability in particle and microorganism concentrations without addressing measurement errors in the concentration values upon which it is based is biased (unless the concentrations are precisely known). Failing to address measurement errors can result in (1) biased analyses if measurement errors cause the concentration values to be inaccurate, (2) extreme concentration values that may not actually occur but that were obtained because the actual concentrations were measured imprecisely, and (3) over-estimated variability in concentration.

This chapter specifically addresses temporal concentration variability (*i.e.* when a series of enumeration data are obtained over time at a specific location), but similar concepts may also apply to spatial variability or to a blend of spatial and temporal variability. The principal difference between this chapter and Chapter 5 is that enumeration data from multiple sampling events are addressed and these data are not assumed to be replicates (*i.e.* they are not representative of the same homogeneous concentration). The models presented herein to address temporal concentration variability are related hierarchically to the models presented in Chapter 5. Temporal concentration variability is particularly important in the case of pathogenic microorganisms in a drinking water supply. Examples in which important regulatory and management decisions are based upon temporally distributed enumerations of microorganisms in water include the United States Environmental Protection Agency's Long Term 2 Enhanced Surface Water Treatment Rule (USEPA, 2006) and quantitative microbial risk assessment (QMRA). This chapter has a particular emphasis upon enumeration of *Cryptosporidium* oocysts and *Giardia* cysts in drinking water, but similar concepts will also apply to the enumeration of other types of microorganisms, to measurement errors in quantitative methods that are not enumeration-based, and to much more diverse applications (*e.g.* the entire scientific field of QMRA must properly address all measurement errors in the data upon which models are based).

The objectives of this chapter are as follows:

- to expand the probabilistic models described in Section 5.1 to account for temporal concentration variability (Section 6.1),
- to evaluate alternative strategies that may enable collection of more informative temporally distributed enumeration data (Section 6.2),
- to apply Bayes' theorem to evaluate the uncertainty in temporal concentration variability given a probabilistic model that addresses the uncertainty in individual concentration estimates due to measurement errors (Section 6.3),
- to discuss the implications of measurement errors upon the Long Term 2 Enhanced Surface Water Treatment Rule (Section 6.4), and
- to evaluate current quantitative microbial risk assessment models that address uncertainty in concentration estimates and to compare these to an approach that is based upon probabilistic modelling of measurement errors and use of Bayes' theorem (Section 6.5).

The chapter is concluded with a discussion of practical implications (Section 6.6).

6.1 MODELLING MEASUREMENT ERRORS IN TEMPORALLY DISTRIBUTED CONCENTRATION DATA

Models that address measurement errors in temporally distributed enumeration-based concentration data describe the relationship between obtained counts and (1) the temporal concentration variability, (2) experimental design factors such as replication within sampling events, the sample volumes, and internal seed recovery data, and (3) analytical errors due to imperfect analytical recovery of the enumeration method. This section addresses expansion of the probabilistic models that were presented in Section 5.1 to include temporal concentration variability (Section 6.1.1) and a discussion of other probabilistic models that have addressed measurement errors in temporally distributed concentration data (Section 6.1.2).

6.1.1 Incorporating Temporal Concentration Variability into Models that Address Measurement Errors

A temporally distributed dataset consists of r sampling events (indexed herein by a subscript i), and each of these sampling events may consist of a single enumeration datum or r_i replicate data (indexed herein by a subscript k). It is assumed herein that the concentration during each sampling event (c_i) is independent and identically distributed. Such a modelling approach assumes that the distribution of concentration over time is stationary (*i.e.* the distribution itself is not changing in time), that sampling events are sufficiently separated in time that temporal autocorrelation can be ignored, and that there is no information about seasonal trends or other possible covariates. These assumptions are discussed in more detail later in this section.

The distribution that is used herein to model temporal concentration variability is the gamma distribution with shape and scale parameters ρ,λ. The gamma distribution has been used extensively in other probabilistic models to describe temporal concentration variability (*e.g.* Teunis *et al.*, 1999; Teunis and Havelaar, 1999; Pouillot, *et al.*, 2004; Petterson *et al.*, 2007; Teunis *et al.*, 2009). Gamma distributions are often considered for continuous, positive, and positively skewed random variables, and the distribution is a mathematically convenient choice for the models presented herein because the gamma distribution is the conjugate prior of the Poisson distribution.

Temporal concentration variability with the beta-Poisson enumeration model

As described in Section 5.1.1, the beta-Poisson enumeration model assumes Poisson-distributed random sampling error, binomially distributed analytical error, and beta-distributed non-constant analytical recovery. It also assumes that analytical recovery is independent of the concentration and sample volume (or at least that the available recovery information is adequately representative of relevant concentrations and sample volumes). The sequence of distributions, along with a hierarchy for multiple sampling events and possible replication within sampling events, is summarized by Equation 6.1: the 'beta-Poisson temporal concentration variability model' that has also been presented by Schmidt and Emelko (2010a). In this equation, n_{ik} is the number of particles collected in a sample of volume V_{ik} from a source with concentration c_i, p_{ik} is the analytical recovery of the enumeration method (with sample-specific parameters for non-constant analytical recovery a_{ik},b_{ik}), and x_{ik} is the number of particles counted in the k^{th} sample of the i^{th} sampling event.

$$\begin{aligned} f\left(\{c_i\},\{n_{ik}\},\{p_{ik}\},\{x_{ik}\} \mid \rho,\lambda,\{V_{ik}\},\{a_{ik}\},\{b_{ik}\}\right) \propto \prod_{i=1}^{r}\left\{\left[\frac{c_i^{\rho-1}e^{-c_i/\lambda}}{\lambda^{\rho}\Gamma(\rho)}\right]\cdot\right. \\ \left.\prod_{k=1}^{r_i}\left[\frac{e^{-c_iV_{ik}}\left(c_iV_{ik}\right)^{n_{ik}}}{n_{ik}!}\right]\left[\binom{n_{ik}}{x_{ik}}p_{ik}^{x_{ik}}\left(1-p_{ik}\right)^{n_{ik}-x_{ik}}\right]\left[\frac{\Gamma\left(a_{ik}+b_{ik}\right)}{\Gamma\left(a_{ik}\right)\Gamma\left(b_{ik}\right)}p_{ik}^{a_{ik}-1}\left(1-p_{ik}\right)^{b_{ik}-1}\right]\right\} \end{aligned} \tag{6.1}$$

As described in Section 5.1, the Poisson distribution used to model random sampling error is based upon an assumption that repeated samples within a single sampling event are independent and that the particles are randomly distributed throughout a source with locally homogeneous concentration. To use the binomial model for analytical error, it must be assumed that each particle has an equal probability of being observed and that analytical recovery cannot exceed 100%. It is assumed that analytical recovery varies randomly (*i.e.* without any non-random or predictable differences) among samples according to a beta distribution. Equation 6.1 is generalized to allow sample-specific recovery parameters a_{ik},b_{ik}, but it is possible that the same parameters could be used for all samples (if, for example, variable water quality and particle concentrations over time do not have a non-random impact upon analytical recovery). If only a fraction (θ_{ik}) of a sample is enumerated, the error due to partial sample analysis can be incorporated into the binomial distribution as per Equation 4.9 in Section 4.2.5.

Temporal concentration variability with internal seed recovery data

If a sample that may possibly contain indigenous particles is seeded with a precisely known number of particles (n^*) that are uniquely identifiable in the final count, it is assumed that each of these seeded particles has an equal probability of being observed, and the probability of recovery (p) is assumed to be equal for seeded and indigenous particles, then Equation 6.1 can be modified to yield Equation 6.2: the ‘internal seed temporal concentration variability model’ that has also been presented by Schmidt and Emelko (2010a). In this equation, x_{ik}^* is the number of seeded particles that were observed in the k^{th} sample of the i^{th} sampling event. The beta distribution describing non-constant analytical recovery is not included in Equation 6.2, but can easily be added. In this model, it is assumed (for simplicity) that p_{ik} is a parameter with a known value. Otherwise, the assumptions for this model are equivalent to those of the beta-Poisson model described above. A hybrid between this model and the beta-Poisson model shown in Equation 6.1 can be used if only some of the samples have internal seed recovery data.

$$\begin{gathered}
f\left(\{c_i\}, \{n_{ik}\}, \{x_{ik}\}, \{x_{ik}^*\} \mid \rho, \lambda, \{V_{ik}\}, \{p_{ik}\}, \{n_{ik}^*\}\right) \propto \prod_{i=1}^{r} \left\{ \left[\frac{c_i^{\rho-1} e^{-c_i/\lambda}}{\lambda^{\rho} \Gamma(\rho)} \right] \cdot \right. \\
\left. \prod_{k=1}^{r_i} \left[\frac{e^{-c_i V_{ik}} (c_i V_{ik})^{n_{ik}}}{n_{ik}!} \right] \left[\binom{n_{ik}}{x_{ik}} p_{ik}^{x_{ik}} (1-p_{ik})^{n_{ik}-x_{ik}} \right] \left[\binom{n_{ik}^*}{x_{ik}^*} p_{ik}^{x_{ik}^*} (1-p_{ik})^{n_{ik}^*-x_{ik}^*} \right] \right\}
\end{gathered} \qquad (6.2)$$

Temporal concentration variability with the negative binomial enumeration model

As described in Section 5.1.2, the negative binomial enumeration model assumes Poisson-distributed random sampling and analytical error and gamma-distributed non-constant analytical recovery. It also assumes that analytical recovery is independent of the concentration and sample volume (or at least that the available recovery information is adequately representative of relevant concentrations and sample volumes). The sequence of distributions, along with a hierarchy for multiple sampling events and possible replication within sampling events, is summarized by Equation 6.3: the ‘negative binomial temporal concentration variability model’. In this equation, x_{ik} is the number of particles observed in a sample of volume V_{ik} from a source with concentration c_i, and p_{ik} is the analytical recovery of the enumeration method (with sample-specific parameters for non-constant analytical recovery α_{ik},β_{ik}). The model is equivalent to the model shown in Equation 6.1 except that analytical recovery is a rate rather than a probability and its variability is described by a gamma distribution. Equation 6.3 is generalized to allow sample-specific recovery parameters α_{ik},β_{ik}, but it is possible that the same parameters could be used for all samples (if, for example, variable water quality and particle concentrations over time do not have a non-random impact upon analytical recovery).

$$\begin{gathered}
f\left(\{c_i\}, \{p_{ik}\}, \{x_{ik}\} \mid \rho, \lambda, \{V_{ik}\}, \{\alpha_{ik}\}, \{\beta_{ik}\}\right) \propto \\
\prod_{i=1}^{r} \left\{ \left[\frac{c_i^{\rho-1} e^{-c_i/\lambda}}{\lambda^{\rho} \Gamma(\rho)} \right] \cdot \prod_{k=1}^{r_i} \left[\frac{e^{-c_i V_{ik} p_{ik}} (c_i V_{ik} p_{ik})^{x_{ik}}}{x_{ik}!} \right] \left[\frac{1}{\beta_{ik}^{\alpha_{ik}} \Gamma(\alpha_{ik})} p_{ik}^{\alpha_{ik}-1} e^{-p_{ik}/\beta_{ik}} \right] \right\}
\end{gathered} \qquad (6.3)$$

Discussion of model assumptions and limitations

> "The main limitations of this model are (1) the appropriateness of the temporal concentration variability distribution, (2) the appropriateness of the Poisson assumption for random sampling error, and (3) the availability of suitable recovery data." – (Schmidt and Emelko, 2010b)

The use of any single distribution to model temporal concentration variability describes the overall variability in concentration over time but ignores temporal correlation and possible variations in the distribution over time due to seasonal trends or other possible covariates. It is probable that microorganism concentrations in a water supply would be temporally correlated because factors such as microorganism loading into the source, dilution, and rates of die-off or decay are likely temporally correlated. Temporal autocorrelation describes the correlation between values as a function of proximity in time; temporally proximate sampling events may have strongly correlated concentrations while widely separated sampling events may have approximately independent concentrations. It is assumed herein that a sequence of concentrations generated using the temporal concentration variability distribution are fully independent and, conversely, that the concentrations estimated from a set of sampling events that are used to estimate the temporal concentration variability distribution are also fully independent. Statistical tools for time series analysis of potentially correlated concentration values must address measurement errors in the concentration estimates to properly evaluate temporal autocorrelation. Concentration values may vary over time due to possible covariates (such as the season, for example), but the modelling approach presented herein describes the overall variability without regard for covariates. It must be assumed that samples are collected in a way that is representative of the overall variability (*e.g.* a sampling schedule that does not over-sample some types of scenarios while under-sampling others). Regression tools that are used to evaluate covariation between microorganism concentrations and other possible covariates must address measurement errors in the concentration estimates to properly evaluate correlations (*e.g.* Crainiceanu *et al.*, 2003) because unaddressed measurement errors can cause excessively variable residuals and the concentration estimates are unlikely to have constant measurement error (*i.e.* the estimates would not be homoscedastic as regression models necessarily assume). The overall temporal concentration variability distribution may also not be stationary (*e.g.* it may be changing over time due to natural or anthropogenic land disturbances). Accordingly, "ongoing sampling or recurrent monitoring programs (*e.g.* USEPA, 2006) are necessary to evaluate possible changes in the pathogen concentration distribution" (Schmidt and Emelko, 2010b). Finally, the gamma distribution used herein may not adequately fit the shape of the overall temporal concentration variability distribution. Other distributions may be more appropriate in general or for case-specific applications, and the models presented herein could be revised accordingly; "evaluation of the fit of the model to the underlying distribution of true concentrations, however, cannot be addressed without first considering the measurement error in the available data" (Schmidt and Emelko, 2010b).

The models that are presented herein assume Poisson-distributed random sampling error. Under- or over-dispersion of microorganism and particle counts with respect to the Poisson model was discussed in Section 5.2. Possible causes for departures from the Poisson model include (1) dependent samples that are falsely regarded as independent replicates, (2) temporal or spatial variability in concentration among samples that are regarded as replicates, (3) imprecisely measured sample volumes, (4) variable analytical recovery of the enumeration method, and (5) non-random particle dispersion in the source. Repeated samples should be independent unless they comprise a large fraction of a source with finite volume. It is assumed herein that the concentration of microorganisms is locally homogeneous within the space and time from which repeated samples (that are regarded as replicates) are drawn. Spatial and temporal heterogeneity among replicates are ignored because it is assumed that multiple samples from the same time and location are representative of the same concentration, and temporal variability among sampling events is addressed in the model. It is assumed that reported sample volumes are reasonably precise; the effect of error in sample volumes upon concentration estimates could be addressed by expanding the variance decomposition analysis in Section 8.2.1. Variability in analytical recovery is specifically addressed in the models presented herein. Alternatives to the Poisson model that may address non-random particle dispersion are not addressed because they would require scale- and case-specific fitting, which would depend upon extensive replication. If the particles are clustered in the source rather than randomly distributed, then the uncertainty in individual concentration estimates will be under-estimated, but the excess variability among sampling events due to clustering (within each sampling event) will be attributed to excess temporal concentration variability.

The models presented herein depend upon a considerable amount of information about analytical recovery. “Evaluating the analytical recovery of an enumeration method may be regarded as less important than evaluating the occurrence of pathogens in drinking water, but enumeration-based concentration estimates are uncalibrated to the actual concentrations that are of real public health significance unless they are adjusted by analytical recovery” (Schmidt and Emelko, 2010b). It is possible to obtain sample-specific recovery data for *Cryptosporidium* oocysts and *Giardia* cysts using internal seeding (*e.g.* Petterson *et al.*, 2007) or to model the variability in analytical recovery among samples using the results of a representative recovery experiment (Schmidt *et al.*, 2010a). In the absence of case-specific recovery information, information from other sources can be used for a preliminary investigation; however, it cannot be assured that the information is suitably representative. A final problem with recovery data (and enumeration data in general) is that the raw data that are necessary to conduct a properly informed statistical analysis are often unavailable. This problem has been raised by Teunis and Havelaar (1999), Roberson *et al.* (2010), and in Section 5.4.4.

6.1.2 Other Probabilistic Models that Address Temporal Concentration Variability

Many probabilistic models for temporal concentration variability are fitted to sets of concentration estimates without consideration of measurement errors. If analytical recovery has been considered at all, it has often been addressed independently of the variability in counts or concentration (Section 6.5.3). This section discusses several other probabilistic models that have addressed measurement errors in temporally distributed concentration data.

Teunis *et al.* (1999) developed ‘probability of passage’ models (Section 5.4.1) to evaluate the reduction in microorganism concentrations by treatment processes when samples were not strictly paired (*e.g.* when there was sub-sampling error in the pre- and post-treatment counts and when the counts were not paired temporally). In these models, random sampling error was addressed using a Poisson distribution and temporal concentration variability was addressed using a gamma distribution. Analytical recovery of the enumeration method was not addressed in any of the ‘probability of passage’ models.

Teunis and Havelaar (1999) developed a probabilistic model similar to the beta-Poisson temporal concentration variability model presented herein. It addressed temporal concentration variability with a gamma distribution, random sampling and analytical error with a Poisson distribution, and non-constant analytical recovery with a beta distribution. Bayes’ theorem was then applied to evaluate the posterior distribution of the temporal concentration variability distribution parameters (ρ,λ). The principal differences between their model and the beta-Poisson temporal concentration variability model presented herein (and its associated Bayesian analysis in Section 6.3.1) are that (1) the nuisance parameters (c, n, and p herein) are integrated out of the model, (2) the model does not account for replication within sampling events, internal seed recovery data, or sample-specific parameters for non-constant analytical recovery, and (3) the Markov Chain Monte Carlo methodology (Gibbs sampling herein) is not presented. The model presented herein enables greater flexibility and is much more accessible to prospective users because the model’s assumptions and limitations are explained and the Gibbs sampling algorithm is clearly presented (Section 6.3.1).

Crainiceanu *et al.* (2003) developed a complicated Bayesian model that addressed measurement errors in recovery and concentration data and that used regression to investigate correlation with possible covariates. The recovery and concentration components are discussed in greater detail in Sections 4.2.8 and 5.1.4 respectively. The model is specific to the United States’ Information Collection Rule *Cryptosporidium* oocyst and *Giardia* cyst datasets (which consisted of samples from many locations that were collected on several occasions and that were enumerated at multiple laboratories). Possible covariates with concentration that were addressed included water quality attributes, sample attributes, and sampling location attributes. The model addressed temporal concentration variability through its relationships with possible covariates that vary in time at a specific site and through random time-site effects. This type of model is particularly useful to investigate correlation between concentration and possible covariates, but it does not quantify temporal concentration variability at a single location in the absence of data for possible covariates.

Pouillot *et al.* (2004) used a negative binomial model to infer the number of oocysts collected in a sample from the number of observed oocysts and the analytical recovery. As described in the beta-Poisson temporal concentration variability model (Section 6.1.1), analytical error is binomially distributed if analytical recovery is a probability. If the probability of recovery associated with a sample is precisely known and there is no prior information on the concentration in the source, a negative binomial model can be used to infer the number of oocysts in the sample that resulted in the obtained count. The negative binomial model used by Pouillot *et al.* (2004) is incorrect, however, for three reasons: (1) it does not work with non-detect samples, (2) the sample-specific probabilities of recovery were not precisely known and were modelled using a beta distribution for non-constant analytical recovery, and (3) the subsequent use of a second negative binomial model (representing

Poisson-distributed random sampling error and gamma-distributed temporal concentration variability) constituted prior information about concentration that was ignored in the first negative binomial model. In effect, the model added variability due to non-constant analytical recovery to counts that already reflected this variability and likely modelled concentrations as being much more variable than what is truly supported by the data. This problem is addressed further in Section 6.5.3. "Had Pouillot *et al.* (2004) used a beta-binomial model for analytical error (which they acknowledged in formulating their negative binomial model for analytical error) in addition to their negative binomial model for random sampling error and temporal concentration variability, they would have arrived at the same model as Teunis and Havelaar (1999) and they would have obtained unbiased maximum likelihood estimates" (Schmidt and Emelko, 2010b). Maximum likelihood estimates, however, do not address uncertainty in the parameters of the distribution describing temporal concentration variability.

Three alternative approaches to analyze temporally distributed (oo)cyst concentration data from a single location were considered by Petterson *et al.* (2007). The first ignored analytical recovery (which essentially assumes that it is consistently 100%), the second addressed variability in analytical recovery estimated from internal seed recovery data corresponding to only a few of the samples, and the third addressed sample-specific evaluation of analytical recovery using internal seed recovery data paired with every sample.

- The first modelling approach addresses variability in concentration using a gamma distribution and random sampling error using a Poisson distribution. These are combined into a negative binomial distribution, and uncertainty in the temporal concentration variability distribution parameters is addressed using Markov Chain Monte Carlo. Unlike the models presented herein, the model assumes fixed sample volumes of 10 L and does not account for possible replication within sampling events. If analytical recovery is not consistently 100%, then this model evaluates the variability in the biased concentration estimates rather than the temporal concentration variability.

- The second approach expands upon the first by evaluating sample-to-sample variability in analytical recovery and the uncertainty therein. It assumes that internal seed recovery data are available for only a few of the samples and that the pairing of these data with counts of indigenous oocysts is ignored. This is equivalent to conducting an independent recovery study, except that the use of internal seeding in relevant samples ensures good representativeness. Bayes' theorem is applied to a beta-binomial model for the recovery data (*i.e.* beta-distributed non-constant analytical recovery and binomially distributed analytical error as described in Section 4.2.2) to evaluate uncertainty in the parameters of the beta distribution that describes variability in recovery. The resulting posterior describing non-constant analytical recovery is integrated with the posterior from the preceding model that describes variability among the biased concentration estimates using a non-Bayesian Monte Carlo approach. The approach assumes that the (oo)cyst counts (or associated concentrations that are not adjusted for analytical recovery) are independent of analytical recovery. Each iteration of the Monte Carlo process consisted of drawing beta and gamma distribution parameters from their respective posteriors, generating 500 gamma-distributed concentrations, and dividing each by a beta-distributed probability of recovery. The resulting unbiased concentration estimates were then fitted with a gamma distribution and the process was repeated many times to evaluate uncertainty in the final gamma distribution fitted to the unbiased concentration estimates. The non-Bayesian approach to integrate information about analytical recovery into the concentration data will result in over-estimated concentration variability. By ignoring the contribution of variable analytical recovery to the variability in the counts, the Bayesian model fitted to the indigenous (oo)cyst count data will over-estimate the variability in concentration. The Monte Carlo process further inflates the over-estimated variability by integrating in variability in analytical recovery. The models presented herein consider temporal variability in actual concentrations, random sampling error, analytical error, and variability in analytical recovery concurrently; application of Bayes' theorem removes the excess variability in the data that is due to measurement errors rather than artificially inflating it. Using Bayes' theorem to evaluate uncertainty in the beta distribution parameters rather than just using point estimates was a novel contribution of the work and has accordingly been addressed in this thesis (*e.g.* Sections 4.4.3, 5.3.5, and 6.3.1).

- The final approach assumes gamma-distributed temporal concentration variability and Poisson-distributed random sampling and analytical error, and incorporates sample-specific internal seed recovery data using a binomial model. The model is similar to the internal seed temporal concentration variability model described herein except that the sample volumes are fixed, there is only one sample per sampling event, and the model cannot address the situation where only a fraction of the (oo)cyst counts are paired with internal seed recovery data. It is not clear how exactly the internal seed recovery data were incorporated into the Bayesian analysis because each sample's probability of recovery is not listed as an unknown parameter of interest and no prior for these recovery values is discussed.

Consideration of measurement errors in the analysis of temporal concentration variability is relatively uncommon. It is imperative, however, to regard enumeration-based concentration estimates as uncertain because of these measurement errors. Many alternative approaches have been proposed to address measurement errors; some properly evaluate different aspects of uncertainty in microbial enumeration, while many over-estimate the variability in concentration because errors are incorporated incorrectly. Assuming that the models presented in Section 6.1.1 appropriately describe measurement errors in the collection of temporally distributed enumeration data, the associated Bayesian analyses described in Section 6.3 provide a framework to evaluate temporal concentration variability with appropriate consideration of measurement errors in the available data. Proper consideration of measurement errors should not be limited to data analysis: it can also be used to strategically enhance the design of future experiments and monitoring programs (Section 6.2) so that more informative data can be collected with the available resources.

6.2 STRATEGIES TO OBTAIN BETTER CONCENTRATION ESTIMATES

Enumeration-based concentration estimates are inherently uncertain because of measurement errors. Strategically targeting the most problematic errors in method development and experimental design will facilitate the collection of more informative data. This section uses Monte Carlo simulation to evaluate the factors that contribute to uncertainty in concentration estimates. This type of analysis could also be used to compare alternative experimental designs.

Herein, the beta-Poisson temporal concentration variability model (along with internal seed recovery data) is used to generate large simulated datasets using Monte Carlo analysis (as per Schmidt and Emelko, 2010a). Three alternative approaches to estimate the particle or microorganism concentration from the data are considered: the count per unit volume, the count per unit volume divided by the mean analytical recovery of the enumeration method, and the count per unit volume divided by a sample-specific internal seed recovery estimate. The first approach is the most common (because analytical recovery is often not considered) and implicitly assumes that the mean analytical recovery of the enumeration method is 100%. The second approach uses the mean analytical recovery estimated from a recovery experiment (*e.g.* Section 4.4.1). The analyses herein assume that the mean analytical recovery is precisely known, but in practice there is always some uncertainty in the estimated average recovery. Sample-specific estimates of internal seed recovery based upon the internal seed recovery data are used in the third approach; specifically, the count per unit volume is divided by the recovery estimate $(x^* + 1)/(n^* + 2)$. This estimate is not unbiased, but is used herein to ensure that recovery estimates of 0% and 100% do not occur.

The factors in the beta-Poisson model that affect the distribution of the counts are the mean and variance of temporal concentration variability (which generally cannot be controlled), the mean and variance of analytical recovery (which can only be improved by selecting or developing better enumeration methodology), and the sample volume (which can be increased, where feasible). The precision of the recovery estimates obtained by internal seeding will be affected by the number of seeded particles (Section 4.3), and this will also affect the precision of the associated concentration estimates. A total of 8 analyses are completed herein, consisting of two temporal concentration variability distributions with four analyses each (which change the variance of analytical recovery, the mean analytical recovery, and the sample volume). The parameters that are used are summarized in Table 6.1.

Each analysis is based upon a total of 10,000 independent simulated counts of indigenous microorganisms. The actual value of concentration that was generated in the simulation process and the three alternative concentration estimates were recorded for each simulation. Each of the four sets of concentrations was then sorted and the associated cumulative relative frequency was plotted. The results of the 8 analyses summarized in Table 6.1 are plotted in Figures 6.1 and 6.2. In these figures, inaccuracy of a concentration estimate is demonstrated by a horizontal shift in the cumulative relative frequency function of the associated estimates relative to the actual concentrations. Imprecision is indicated by a cumulative relative frequency function that is broader than that of the actual

Table 6.1: Parameters used in Monte Carlo simulations

Analysis	**Concentration** (microorganisms/L)	**Recovery** (%)	**Volume** (L)	**Seed Dose** (microorganisms)	**Figure**
1	$\rho = 1$, $\lambda = 10$ ($\mu_c = 10$, $\sigma_c = 10$)	$a = 3$, $b = 7$ ($\mu_p = 30.0$, $\sigma_p = 13.8$)	10	100	Figure 6.1a
2		$a = 30$, $b = 70$ ($\mu_p = 30.0$, $\sigma_p = 4.56$)	10	100	Figure 6.1b
3		$a = 7$, $b = 3$ ($\mu_p = 70.0$, $\sigma_p = 13.8$)	10	100	Figure 6.1c
4		$a = 3$, $b = 7$ ($\mu_p = 30.0$, $\sigma_p = 13.8$)	100	100	Figure 6.1d
5	$\rho = 1$, $\lambda = 10$ ($\mu_c = 10$, $\sigma_c = 1$)	$a = 3$, $b = 7$ ($\mu_p = 30.0$, $\sigma_p = 13.8$)	10	100	Figure 6.2a
6		$a = 30$, $b = 70$ ($\mu_p = 30.0$, $\sigma_p = 4.56$)	10	100	Figure 6.2b
7		$a = 7$, $b = 3$ ($\mu_p = 70.0$, $\sigma_p = 13.8$)	10	100	Figure 6.2c
8		$a = 3$, $b = 7$ ($\mu_p = 30.0$, $\sigma_p = 13.8$)	100	100	Figure 6.2d

concentrations (because the concentration estimates are more variable than concentration itself). Figures 6.1 and 6.2 are used below to evaluate the improvement in accuracy associated with reporting recovery-adjusted concentration estimates and the improvement in precision that may be obtained by decreasing the variability in analytical recovery, increasing the mean analytical recovery, or increasing the sample volume.

Improving the accuracy of concentration estimates by factoring in analytical recovery

Figures 6.1 and 6.2 clearly show that concentration estimates that ignore the analytical recovery of the enumeration method are biased. In many applications, it is the actual microorganism concentration that is of scientific or regulatory interest rather than the number of microorganisms observed in a particular sample volume. In most of these examples, the mean analytical recovery is 30%, which corresponds to a bias of approximately 0.5-log if analytical recovery is ignored. When the mean analytical recovery is increased to 70%, the bias associated with ignoring analytical recovery in concentration estimates decreases to 0.15-log on average (Figures 6.1c and 6.2c). This bias vanishes as the mean analytical recovery approaches 100%. Petterson *et al.* (2007) described the relationship between this bias and analytical recovery by saying that "the consequences of ignoring imperfect detection increase as the recovery worsens". Concentration estimates can be calibrated to actual concentrations using the mean analytical recovery of the enumeration method or by obtaining sample-specific recovery estimates such as those obtained by internal seeding. Such calibration will make the concentration estimates more accurate, but will not necessarily make them more precise.

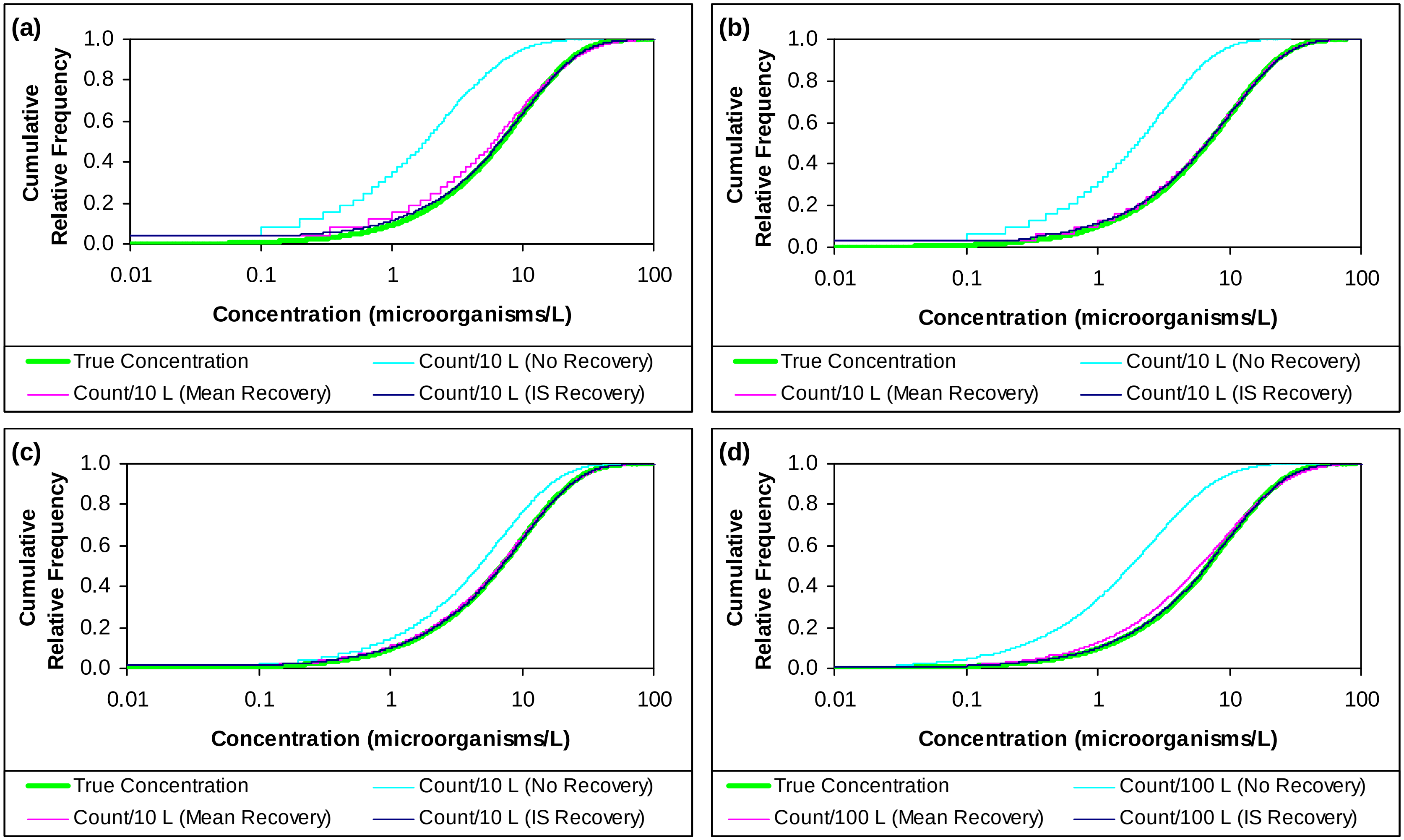


Figure 6.1: Factors that affect concentration estimates when temporal concentration variability is high

These analyses are based upon 10,000 simulations using ρ = 1, λ = 10 (μ_c = 10, σ_c = 10 in microorganisms/L). The panels represent (a) a base scenario against which other panels are compared [a = 3, b = 7, V = 10, n^* = 100], (b) a scenario with less variable analytical recovery [a = 30, b = 70, V = 10, n^* = 100], (c) a scenario with higher mean recovery [a = 7, b = 3, V = 10, n^* = 100], and (d) a scenario with higher sample volume [a = 3, b = 7, V = 100, n^* = 100]. 'No Recovery', 'Mean Recovery' and 'IS Recovery' describe what type of adjustment for recovery is used.

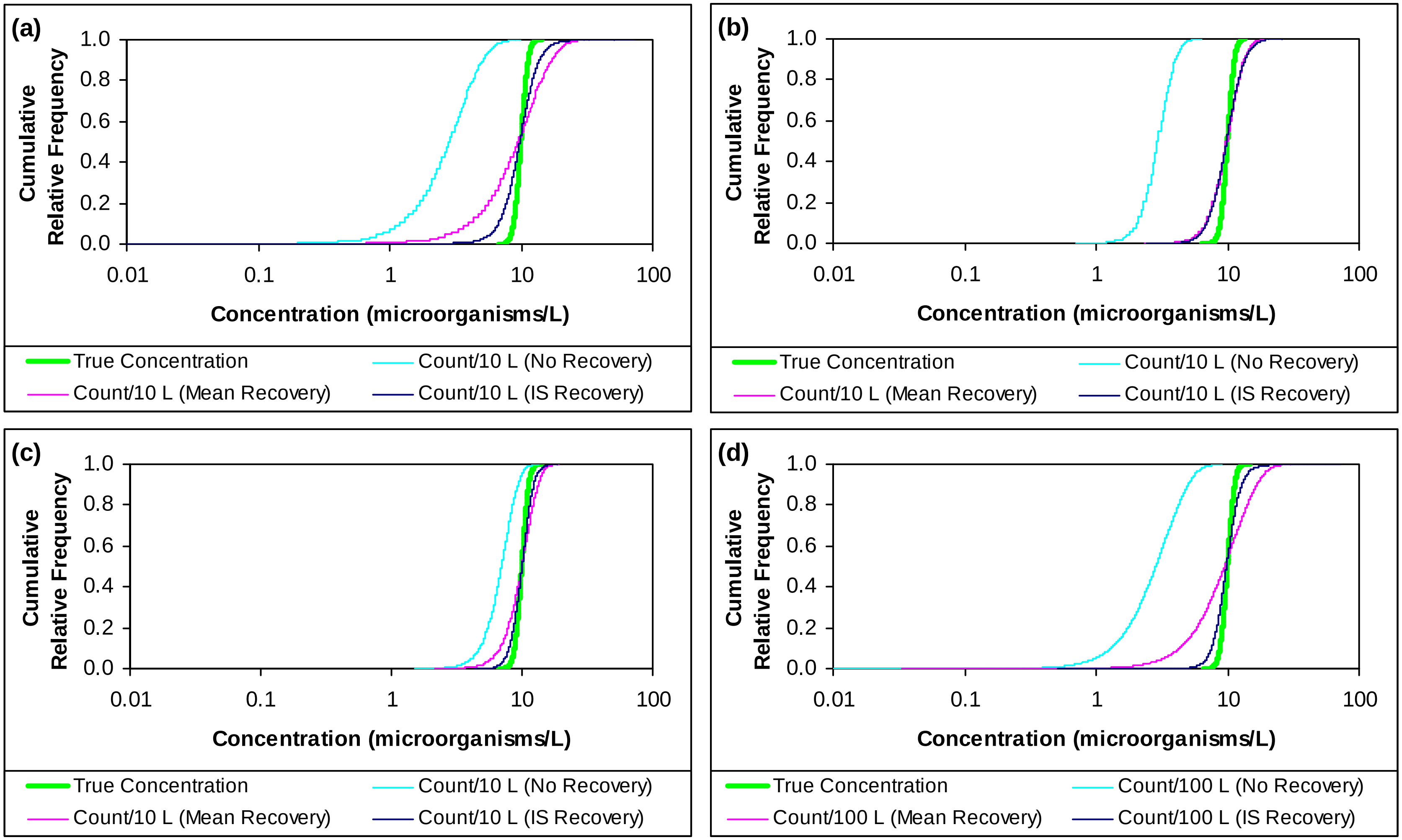


Figure 6.2: Factors that affect concentration estimates when temporal concentration variability is low

These analyses are based upon 10,000 simulations using $\rho = 10$, $\lambda = 1$ ($\mu_c = 10$, $\sigma_c = 1$ in microorganisms/L). The panels represent (a) a base scenario against which other panels are compared [$a = 3$, $b = 7$, $V = 10$, $n^* = 100$], (b) a scenario with less variable analytical recovery [$a = 30$, $b = 70$, $V = 10$, $n^* = 100$], (c) a scenario with higher mean recovery [$a = 7$, $b = 3$, $V = 10$, $n^* = 100$], and (d) a scenario with higher sample volume [$a = 3$, $b = 7$, $V = 100$, $n^* = 100$]. ‘No Recovery’, ‘Mean Recovery’ and ‘IS Recovery’ describe what type of adjustment for recovery is used.

Improving the precision of concentration estimates by reducing the variability of analytical recovery

Measurement error in enumeration methods can be reduced by increasing mean analytical recovery or by decreasing the variability in analytical recovery. Stricter control of methodology may lead to some improvement in analytical recovery, but recovery can generally only be improved by development and selection of superior methodology. Before investing resources into method development or more complicated methodology, it is important to consider whether improvements will actually lead to better results. Figures 6.1b and 6.2b demonstrate the effect of reduced variability in analytical recovery relative to Figures 6.1a and 6.2a respectively. In Figure 6.1, the percentage of non-detect samples fell slightly from 4.34% to 3.46% (which indicates that variable analytical recovery is a contributing factor to poor sensitivity) and the calibrated concentration estimates became slightly more precise when the variability in analytical recovery was reduced. In this particular scenario, neither reducing the variability of analytical recovery nor the use of sample-specific internal seed recovery estimates substantially improved the precision of the concentration estimates. In Figure 6.2, however, reducing the variability in analytical recovery substantially improved the precision of the concentration estimates that were adjusted by mean analytical recovery. This is because the mean recovery is a better estimate of the sample-specific recovery when the variability in analytical recovery is reduced. The precision of the concentration estimates that were adjusted by internal seed recovery estimates was not improved, however, because this approach already accounted for the variability in analytical recovery. Use of sample-specific recovery estimates in Figure 6.2a enabled collection of more precise recovery estimates, but the added complexity of internal seeding in Figure 6.2b yields little improvement in precision. In each scenario, further analyses (or variance decomposition as discussed in Section 8.2) would be required to evaluate which of the following would best reduce the remaining imprecision of the recovery-adjusted concentration estimates: increasing the sample volume, further reducing the variability in analytical recovery, or increasing the precision of the sample-specific recovery estimates. Comparison of Figures 6.1 and 6.2 shows that reducing the variability in analytical recovery has greater effect when concentration is less variable. This is proven by variance decomposition in Section 8.2.2. Reducing a source of variability that is a small contributor to overall variability will have minimal effect.

Improving the precision of concentration estimates by increasing the mean analytical recovery

Increasing the mean analytical recovery of the enumeration method reduces the analytical error by increasing the number of microorganisms or particles in each sample that are actually observed (on average). This will in turn result in improved sensitivity: 4.34% of samples in Figure 6.1a were non-detects, and this was reduced to 1.36% in Figure 6.1c by increasing the mean analytical recovery from 30% to 70%. The reduction in analytical error also results in the concentration estimates that do not account for analytical recovery becoming less biased and makes each type of concentration estimate much more precise (*e.g.* Figures 6.1c and 6.2c).

Improving the precision of concentration estimates by increasing sample volumes

Increasing the volume of samples (where feasible and if it does not adversely affect the analytical recovery of the enumeration method) will reduce random sampling error. Figures 6.1d and 6.2d demonstrate the effects of increased sample volume relative to Figures 6.1a and 6.2a respectively. In Figure 6.1, increasing the sample volumes from 10 L to 100 L reduced the percentage of non-detect samples from 4.34% to 0.0047% and improved the overall precision of the recovery-adjusted concentration estimates. Of the three alternatives considered, this increase in sample volume yielded the greatest improvement in sensitivity. In this particular scenario, the concentration estimates that are adjusted by the mean analytical recovery are still somewhat imprecise while the concentration estimates that are adjusted by sample-specific recovery estimates are quite precise. This suggests that the remaining measurement error in the ‘Mean Recovery’ concentration estimates is due to variability in analytical recovery (which is suitably addressed by the ‘IS Recovery’ concentration estimates). In Figure 6.2, the precision of both types of recovery-adjusted concentration estimates is modestly improved by increasing the sample volume, but the concentration estimates are still much more variable than concentration itself. The concentration estimates that are calibrated using sample-specific recovery estimates are much more precise than the estimates that are adjusted by mean analytical recovery in Figures 6.2a and 6.2d, which shows that much of the error in the latter type of estimates is due to variability in analytical recovery. In general, increasing sample volumes will reduce the overall variability in concentration estimates (or uncertainty in individual concentration estimates as shown in Section 5.6.2) if random sampling error is a substantial source of variability; otherwise, the increased complexity that may be associated with larger sample volumes may not be worthwhile. Increasing the sample volumes had a greater impact in Figure 6.2 than Figure 6.1 because less of the overall variability was due to variability in concentration itself. Increasing the sample volume and then enumerating only a portion of the sample does not improve the precision of concentration estimates unless the effective volume (Section 5.1.3) is increased. Variance decomposition is used to evaluate when larger sample volumes enable more precise concentration estimates in Section 8.2.2.

Improving the precision of concentration estimates by obtaining internal seed recovery estimates

If analytical recovery is highly variable (particularly if it varies non-randomly among samples so that the variability cannot be described appropriately by a distribution), then adjusting concentration estimates by sample-specific recovery estimates can improve precision. Sample-specific recovery estimates can be obtained through the internal seeding process that was modelled in Section 6.1.1. Compared to concentration estimates that are adjusted by the mean analytical recovery, the concentration estimates that are based upon internal seed recovery estimates are slightly more precise in Figures 6.1a and 6.1d, and substantially more precise in Figures 6.2a, 6.2c, and 6.2d. If, however, analytical recovery does not vary substantially, these recovery estimates may not be much more informative than the mean analytical recovery of the enumeration method (*e.g.* Figures 6.1b and 6.2b). It is not only the variability in analytical recovery that affects the relative value of internal seed recovery data, but the proportion of the overall variability that is attributed to variability in analytical recovery. When the variability in concentration itself was greater (Figure 6.1), the relative value of concentration estimates that are adjusted by sample-specific recovery estimates (as compared to just being adjusted by mean recovery) is less (*e.g.* compared to Figure 6.2). "Even if sample-specific recovery estimates do not yield more accurate concentration estimates, internal seeding of a random selection of samples may be a more convenient and representative approach to evaluate recovery than conducting a fully independent recovery experiment" (Schmidt and Emelko, 2010a). "Internal spiking of labeled [oo]cysts into environmental samples overcomes many limitations experienced by isolated laboratory recovery trials, in particular the need to replicate the specific water matrix of the environmental samples" (Petterson *et al.*, 2007).

A problem with internal seed recovery estimates (and recovery estimates in general) is that they are imprecise as demonstrated in Section 4.3. It is plausible that there may be scenarios in which the imprecision of internal seed recovery estimates may increase the imprecision of the associated concentration estimates relative to concentration estimates that are adjusted by mean analytical recovery. The types of Monte Carlo simulations shown in this section can be used to determine whether or not the added complexity of using internal seeding to obtain sample-specific recovery estimates is worthwhile.

6.3 QUANTIFYING TEMPORAL CONCENTRATION VARIABILITY USING BAYES' THEOREM

The models presented in Section 6.1.1 describe the distribution of possible counts ($\{x_{ik}\}$, and $\{x_{ik}^*\}$ where applicable) given many parameters that are assumed to be known ($\rho,\lambda,\{V_{ik}\},\{a_{ik},b_{ik}\}$ or $\{a_{ik},b_{ik}\}$, and $\{n_{ik}^*\}$ where applicable). The sampling event concentrations $\{c_i\}$ are nuisance parameters because they are not actually known. The numbers of indigenous particles contained in each sample ($\{n_{ik})\}$ in the beta-Poisson temporal concentration variability model is also a nuisance parameter. In the beta-Poisson and negative binomial temporal concentration variability models, the analytical recovery values ($\{p_{ik}\}$) are nuisance parameters. The internal seed temporal concentration variability model, however, treats analytical recovery values as known parameters. Such a model can be used to simulate count data (*e.g.* Sections 6.2, 6.3.2, and 6.5.2) or as the basis to properly evaluate unknown model parameters given available data.

When evaluation of unknown parameters given experimental data is the objective, ρ and λ are typically the unknown parameters of interest and the nuisance parameters also need to be evaluated. Bayes' theorem can be applied to these probabilistic models to obtain a posterior that describes what the analyst is entitled to believe about the temporal concentration variability distribution parameters (ρ,λ) given the model, the available data, and a specific prior. The prior for these parameters is generically represented by $g(\rho,\lambda)$ herein. Priors do not need to be specified for nuisance parameters because the prior information about these parameters is fully specified by the probabilistic model. Relatively uninformative semi-infinite uniform priors for ρ and λ (on the parameter space $\rho > 0$, $\lambda > 0$) are used herein for demonstration purposes. These priors are improper (because their respective integrals are infinite) and may potentially lead to an improper posterior. Further research is needed to evaluate the propriety of the resulting posterior in each of the models. In the internal seed temporal concentration variability model, the analytical recovery values are also unknown parameters of interest (unless the model is modified to include a non-constant analytical recovery distribution for each sample or the recovery values are precisely known by the analyst). Therefore priors must also be specified for $\{p_{ik}\}$ in the internal seed model. Herein, the relatively uninformative uniform prior on the interval $0 < p_{ik} < 1$ is used. Incorporation of beta-distributed informative priors is also discussed.

Evaluating these posteriors by integration is computationally intractable because several levels of numerical integration would be necessary. If it is assumed that the posterior is proper, then Gibbs sampling (Section 3.2.2) is a much easier alternative to evaluate complicated posterior distributions. Development of Gibbs sampling algorithms for the beta-Poisson and internal seed temporal concentration variability models is discussed in Section 6.3.1. Additional details for these models, and also for the negative binomial temporal concentration variability model, is provided in Appendix B.5. Section 6.3.2 provides examples of the Gibbs sampling approach along with discussion of mixing and convergence.

6.3.1 Evaluating Posterior Distributions Using Gibbs Sampling

Application of Bayes' theorem to the beta-Poisson and internal seed temporal concentration variability models yields the relationships presented in Equations 6.4 and 6.5.

$$f\left(\rho, \lambda, \{c_i\}, \{n_{ik}\}, \{p_{ik}\} \mid \{x_{ik}\}\right) \propto g(\rho, \lambda) \cdot f\left(\{c_i\}, \{n_{ik}\}, \{p_{ik}\}, \{x_{ik}\} \mid \rho, \lambda\right) \tag{6.4}$$

$$f\left(\rho, \lambda, \{c_i\}, \{n_{ik}\}, \{p_{ik}\} \mid \{x_{ik}\}, \{x_{ik}^*\}\right) \propto g\left(\rho, \lambda, \{p_{ik}\}\right) \cdot f\left(\{c_i\}, \{n_{ik}\}, \{x_{ik}\}, \{x_{ik}^*\} \mid \rho, \lambda, \{p_{ik}\}\right) \tag{6.5}$$

If the posterior is proper, then Gibbs sampling can be used to generate a Markov chain of the vector of parameters that describes the posterior distribution. Gibbs sampling requires initial values of the unknown parameters, an iterative loop in which a new value of each unknown parameter is generated from its conditional posterior distribution (given the most recent value of each parameter upon which it depends), and methods to sample from each of the conditional posterior distributions. The conditional posterior distributions for the beta-Poisson temporal concentration variability model are summarized by Equations 6.6 to 6.10. Most of the conditional posteriors of the internal seed temporal concentration variability model are the same because the model is identical except that non-constant analytical recovery beta distributions are replaced with internal seed binomial distributions and a prior for the recovery values $\{p_{ik}\}$. Equation 6.11 is the conditional posterior distribution for p_{ik} assuming a uniform prior for recovery. Equation 6.11 can be regarded as a special case of Equation 6.10 in which $a_{ik} = x_{ik}^* + 1$, $b_{ik} = n_{ik}^* - x_{ik}^* + 1$ (Schmidt and Emelko, 2010a). Accordingly, the information from internal seed data can be incorporated into the beta-Poisson temporal concentration variability model as sample-specific recovery parameters a_{ik},b_{ik}. This is particularly convenient if only some of the enumeration data are paired with internal seed recovery data. An informative beta-distributed prior for recovery (*e.g.* to represent non-constant analytical recovery), p_{ik} ~ BETA(a_{ik}^*,b_{ik}^*), would also result in a beta-distributed conditional posterior for p_{ik}. In this case, Equation 6.10 can be used with the substitution $a_{ik} = x_{ik}^* + a_{ik}^*$, $b_{ik} = n_{ik}^* - x_{ik}^* + b_{ik}^*$ (Schmidt and Emelko, 2010a).

$$Dfc(\rho) \propto g(\rho \mid \lambda) \cdot (\Gamma(\rho))^{-r} \left(\frac{1}{\lambda} \prod_{i=1}^{r} c_i \right)^{\rho} \tag{6.6}$$

$$Dfc(\lambda) \propto g(\lambda \mid \rho) \cdot \lambda^{-r\rho} \cdot e^{-\frac{1}{\lambda}\sum c_i} \tag{6.7}$$

$$Dfc(c_i) \propto c_i^{\alpha-1+\sum n_{ik}} \cdot e^{-c_i\left(\frac{1}{\beta}+\sum V_{ik}\right)} \to c_i \left(\frac{1}{\beta} + \sum_{k=1}^{r_i} V_{ik} \right) \sim GAMMA\left(\alpha + \sum_{k=1}^{r_i} n_{ik}, 1 \right) \tag{6.8}$$

$$Dfc(n_{ik}) \propto \frac{(c_i V_{ik} (1 - p_{ik}))^{n_{ik}}}{(n_{ik} - x_{ik})!} \to n_{ik} - x_{ik} \sim POISSON(c_i V_{ik} (1 - p_{ik})) \tag{6.9}$$

$$Dfc(p_{ik}) \propto p_{ik}^{x_{ik}+a_{ik}-1} (1 - p_{ik})^{n_{ik}-x_{ik}+b_{ik}-1} \to p_{ik} \sim BETA(x_{ik} + a_{ik}, n_{ik} - x_{ik} + b_{ik}) \tag{6.10}$$

$$\begin{aligned} &Dfc(p_{ik}) \propto p_{ik}^{x_{ik}+x_{ik}^*} (1 - p_{ik})^{n_{ik}+n_{ik}^*-x_{ik}-x_{ik}^*} \\ &\to p_{ik} \sim BETA(x_{ik} + x_{ik}^* + 1, n_{ik} + n_{ik}^* - x_{ik} - x_{ik}^* + 1) \end{aligned} \tag{6.11}$$

The initial value of each parameter that is used in Gibbs sampling can be any feasible value, but a smaller burn-in is possible if reasonably probable initial parameter values are used. The initial values used herein are $p_{ik} = a_{ik} / (a_{ik} + b_{ik})$ (with the substitution described above for internal seeding) and $c_i = \Sigma x_{ik} / \Sigma(V_{ik} p_{ik})$ (with Σx_{ik} set to a value of 0.5 if the sum of the counts is zero). Initial values for ρ and λ were obtained using the method of moments with the initial values of $\{c_i\}$. The sequential order of the Gibbs sampling algorithm is summarized by Figure 6.3.

Figure 6.3 also shows how to generate a sequence of concentration values $\{c^*\}$ that is representative of the posterior predictive distribution (Gelman *et al.*, 2004) of concentration and how to incorporate uncertainty in beta distribution parameters (a_{ik},b_{ik}) that are estimated from the results of a recovery experiment. The posterior predictive distribution of concentration represents the variability in concentration with consideration of measurement errors in the enumeration data as well as uncertainty in the temporal concentration variability distribution parameters ρ,λ. Point estimates of the parameters a_{ik},b_{ik} can be obtained from the results of recovery experiments using maximum likelihood estimation (*e.g.* Teunis *et al.*, 1999; Schmidt *et al.*, 2010a). These estimates, however, are uncertain as discussed in Section 4.4.3. Uncertainty in these parameters can be incorporated into Gibbs sampling using a Monte Carlo process in which a random pair of parameter values is supplied from their joint posterior distribution for each Gibbs sampling iteration. A sequence of such parameter pairs could be generated using the recovery model Gibbs sampling process described in Section 4.4.3, and these can subsequently be inserted into the present Gibbs sampling algorithm in randomized order. Incorporation of uncertainty in the beta distribution parameters was addressed in the analysis of replicate concentration data in Section 5.3.5, but is not addressed herein. Petterson *et al.* (2007) also used a Markov Chain Monte Carlo process to address uncertainty in these parameters.

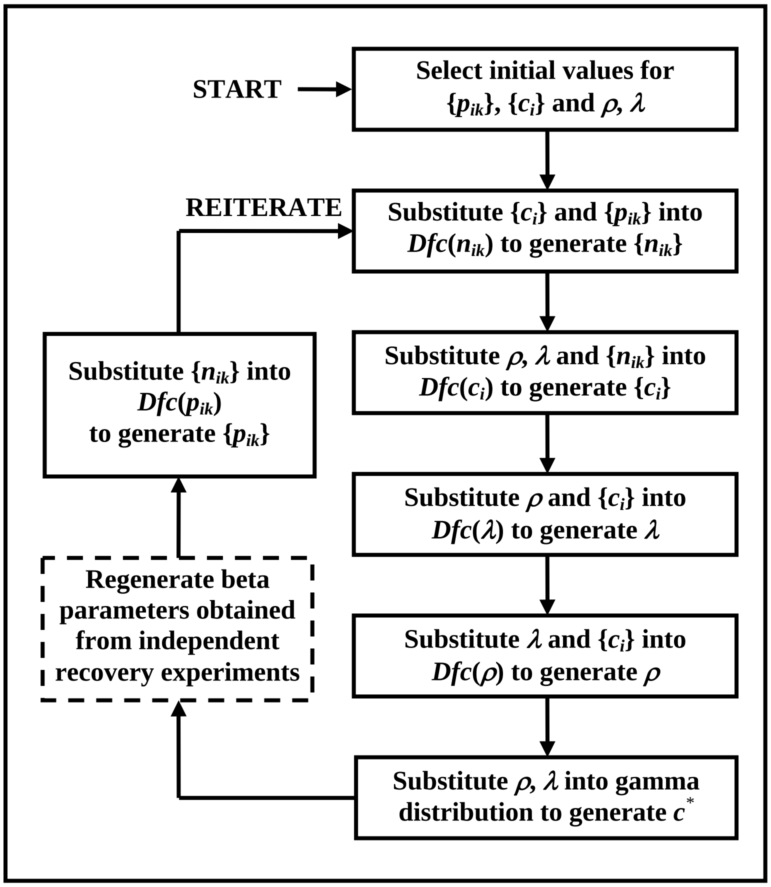

START
Select initial values for
{p_{ik}}, {c_i} and ρ, λ
REITERATE
Substitute {c_i} and {p_{ik}} into
$Dfc(n_{ik})$ to generate {n_{ik}}
Substitute {n_{ik}} into
$Dfc(p_{ik})$
to generate {p_{ik}}
Substitute ρ, λ and {n_{ik}} into
$Dfc(c_i)$ to generate {c_i}
Substitute ρ and {c_i} into
$Dfc(\lambda)$ to generate λ
Regenerate beta
parameters obtained
from independent
recovery experiments
Substitute λ and {c_i} into
$Dfc(\rho)$ to generate ρ
Substitute ρ, λ into gamma
distribution to generate c^*

Figure 6.3: Gibbs sampling algorithm for beta-Poisson concentration variability
(modified from Schmidt and Emelko, 2010a)
The Gibbs sampling algorithm begins with a burn-in and ends after a user-specified number of iterations. The sequence of c^* concentration values is representative of the posterior predictive distribution of concentration. Uncertainty in the parameters of the beta distribution describing non-constant analytical recovery can be addressed by using new estimated parameter values from a posterior distribution (*e.g.* using Gibbs sampling as discussed in Section 4.4.3).

With the exception of Equation 6.6, generating new parameter values using the conditional posterior distributions is relatively simple. Acceptance-rejection sampling is used to generate values of ρ because Equation 6.6 cannot be arranged in the form of a distribution for which algorithms to generate random numbers are readily available. The process that is used has been described in Schmidt and Emelko (2010a) and is described for a different conditional posterior distribution in Section 4.4.3. Proof that this conditional posterior distribution is unimodal with the uniform prior used herein is provided in Appendix A. If $g(\lambda|\rho)$ is either an improper semi-infinite uniform prior or a gamma distribution, then the conditional posterior for λ^{-1} is gamma-distributed. Given the uniform prior used herein, Equation 6.7 can be rearranged as $\lambda^{-1} \cdot \Sigma c_i \sim GAMMA(r\rho + 1, 1)$. This Gibbs sampling algorithm is used in Section 6.3.2, and the results are used in Section 6.5.3.

6.3.2 Using Gibbs Sampling in the Analysis of an Example Dataset

This section addresses the analysis of an example dataset that was presented in Schmidt and Emelko (2010b) using the preceding Gibbs sampling approach. The dataset (Appendix D.4) was simulated using the internal seed temporal concentration variability model (Equation 6.2) with beta-distributed non-constant analytical recovery. A total of 24 sampling events (each with only one sample) were simulated using the following parameter values: ρ = 0.22, λ = 0.36, and (V = 100 L, n^* = 100 oocysts, a = 2, b = 3) for all sampling events. The mean and standard deviation of concentration are 0.0792 oocysts/L and 0.1689 oocysts/L respectively, and the mean and standard deviation of analytical recovery are 0.4 and 0.2 respectively. These particular beta distribution parameter values were used by Jaidi *et al.* (2009) to represent typical analytical recovery of *Cryptosporidium* oocysts in surface water by Method 1623 (USEPA, 2005b). The values of ρ and λ were chosen to yield comparable summary statistics to those presented for a set of *Cryptosporidium* counts in temporally distributed 100 L surface water samples (Jaidi *et al.*, 2009). The mean and standard deviation of the counts per unit volume (*i.e.* without consideration of analytical recovery) in the simulated data are 3.0 and 8.7 oocysts/100 L respectively, and 8 of the 24 data are non-detects.

The dataset is analysed using the beta-Poisson temporal concentration variability model and Gibbs sampling in each of two ways: (1) using a = 2, b = 3 (to represent the case in which internal seed recovery data are unavailable) and (2) using $a_{ik} = x_{ik}^* + 1$, $b_{ik} = n_{ik}^* - x_{ik}^* + 1$ (to represent the case in which internal seed recovery data are available and a relatively uninformative uniform prior is used for analytical recovery). A third Gibbs sampling sequence using a = 2, b = 3 and the extreme initial parameter values ρ = 1000, λ = 1000 is shown to illustrate good convergence. In each case, a burn-in of 1000 iterations was followed by 30,000 iterations that were used to evaluate the posterior distribution. The resulting sequences of ρ and λ are plotted in Figure 6.4. Figures 6.4a and 6.4b show very good mixing and suggest that reasonable convergence might be attained after 30,000 iterations (although more iterations would be needed to obtain very precise posterior probabilities, especially in the tails of the distribution). Figure 6.4c shows good convergence from extreme initial parameter values, albeit slowly. In some other runs (not shown), the sequence did not converge from extreme initial parameter values (especially using the beta parameters based upon internal seed recovery data). It is plausible that the improper semi-infinite uniform priors for ρ and λ could yield an improper posterior, and further investigation of relatively uninformative priors is recommended. Posterior predictive distributions of concentration based upon the first two Gibbs sampling sequences are discussed in Section 6.5.4.

6.4 CASE STUDY: CONSIDERATION OF MEASUREMENT ERRORS IN THE LT2ESWTR

Determination of the level of treatment that is required to consistently ensure adequately safe drinking water (with respect to pathogens) requires knowledge of the health impacts of various quantities of pathogens in the treated water and knowledge of the quantities of pathogens that are present in the

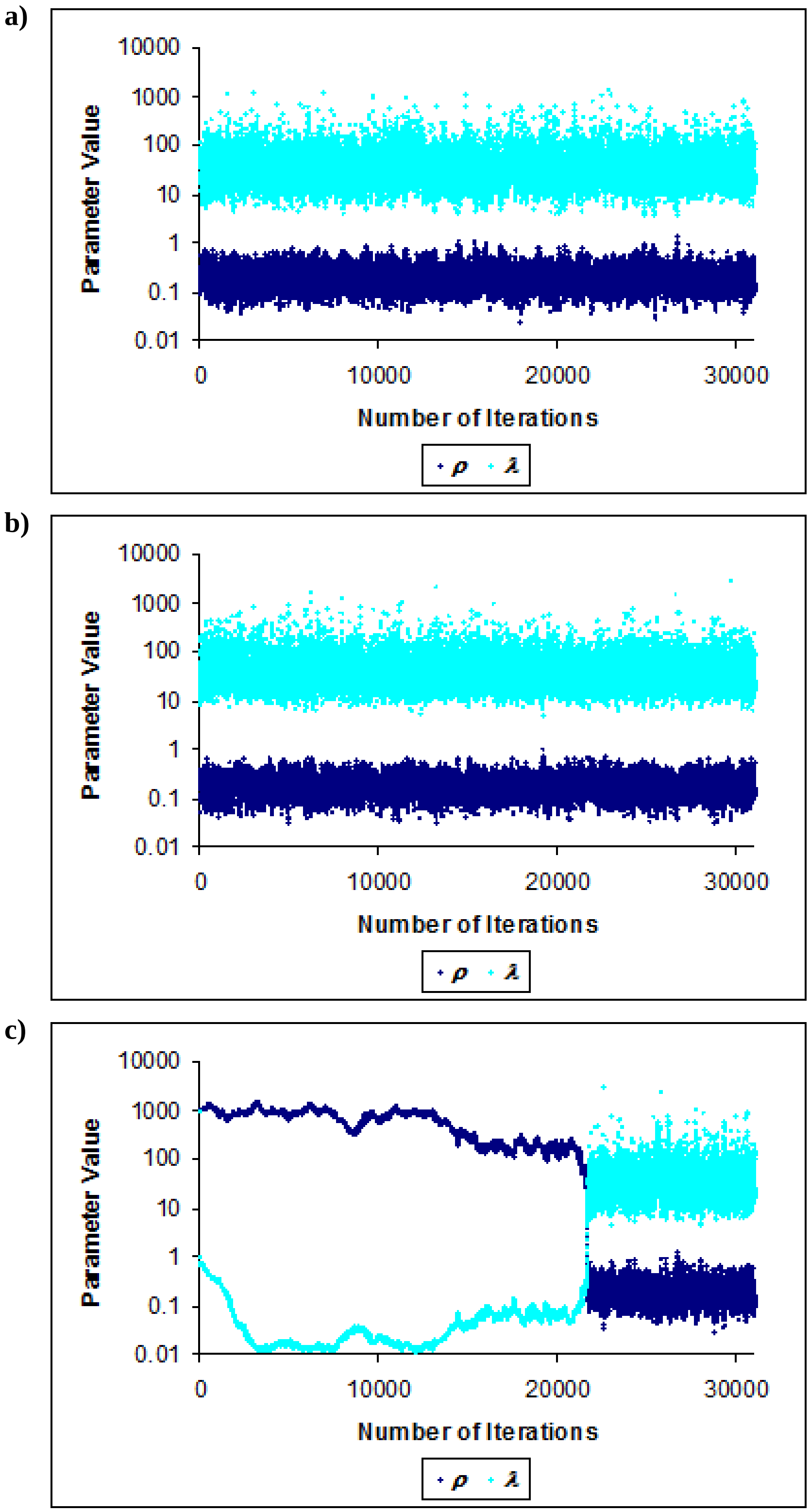


Figure 6.4: Gibbs sampling associated with temporal concentration variability models

These Gibbs sampling results are based on the Table D.4 enumeration data using the beta-Poisson temporal concentration variability model with (a) $a = 2$, $b = 3$, (b) beta parameters based upon the internal seed recovery data, and (c) $a = 2$, $b = 3$ starting with $\rho = 1000$, $\lambda = 1000$.

untreated water (*e.g.* Haas *et al.*, 1996; Medema *et al.*, 2003). The United States' Long Term 2 Enhanced Surface Water Treatment Rule (USEPA, 2006) is a regulation that requires specific levels of treatment (with respect to *Cryptosporidium* oocysts) depending upon the average oocyst concentration in the raw water. For example, public water systems that use surface water (or ground water under direct influence of surface water) and that serve more than 10,000 people were required to monitor source water *Cryptosporidium* concentrations at least monthly for a period of two years. The average concentration was then used to determine in which of four bins of prescribed *Cryptosporidium* treatment requirements the public water system fell.

> "When determining the *Cryptosporidium* bin concentration, [public water systems] must calculate individual sample concentrations as the total number of oocysts counted, divided by the volume assayed. In samples where no oocysts are detected, the result is assigned a value of zero for the purpose of calculating the bin concentration. Sample analysis results are not adjusted for analytical method recovery or the percent of *Cryptosporidium* oocysts that are infectious." – (USEPA, 2006)

Analytical recovery and infectivity were not addressed in the regulation because "these two factors are offsetting, in that oocyst counts not adjusted for recovery tend to underestimate the true concentration, while the total oocyst count typically overestimates the infectious concentration that presents a health risk" and because "it is not possible to establish a precise value for method recovery or the fraction of oocysts that are infectious, [but] available data suggest that these parameters may be of similar magnitude" (USEPA, 2006). The assumption that analytical recovery and infectivity are counter-balancing factors is discussed further in Section 6.5.2.

The regulation addressed bin misclassification error, which is "due to the limited number and volume of samples that can be analyzed, imperfect method recovery, and variability in *Cryptosporidium* occurrence" because it could lead to public water systems "not providing an adequate level of treatment" or "incurring additional costs for unnecessary treatment" (USEPA, 2006). The sampling requirements (*i.e.* the number of samples) were designed to ensure acceptably low levels of bin misclassification error. Emelko *et al.* (2008) proposed that the regulation should address uncertainty in concentration estimates due to measurement errors. The concentration estimates that are used to calculate the average concentration are potentially inaccurate (unless analytical recovery and infectivity are equal) and are certainly imprecise (*i.e.* measurement errors make concentration estimates uncertain even if they are unbiased). Schmidt and Emelko (2010a) proposed that "analytical recovery information should be incorporated into drinking water regulations so that the concentration metric that is used to evaluate compliance (or the need for mandatory mitigation) is equally stringent for all regulated water systems". The problem with the present regulation is that "utilities that obtain higher analytical recovery (due to raw water quality, laboratory, or method used) are more likely to observe any (oo)cysts that are present and to incur mandatory treatment enhancement, while utilities that obtain lower analytical recovery are less likely to observe present (oo)cysts and may accordingly under-predict health risks" and that "failing to account for recovery may indirectly promote use of poor methodology by penalizing those who typically obtain higher recovery (*i.e.* observe a higher fraction of the parasites that are present)" (Schmidt and

Emelko, 2010a). Olstadt *et al.* (2007) and Bennear *et al.* (2009), for example, have presented ways in which coliform monitoring results can potentially be manipulated: selecting inferior methodology so that fewer coliform bacteria are detected or collecting additional samples until a favourable result is obtained ('sampling out').

Knowledge of measurement errors and analytical recovery should be addressed in future monitoring programs, such as the upcoming second round of monitoring under the LT2ESWTR (USEPA, 2006), to ensure that decisions are made with proper regard for measurement errors in the data and uncertainty in estimated quantities such as the average concentration. If analytical recovery information is collected along with the temporally distributed oocyst occurrence data, then the Bayesian methods presented herein can be used to evaluate uncertainty in the mean concentration (without consideration of infectivity) and also to evaluate the variability in concentration. Such analyses are recommended because (1) they would ensure that regulatory decisions are equally stringent for all public water systems regardless of the analytical recovery that each system's laboratory typically achieves, (2) each system could make management decisions with appropriate knowledge of uncertainty in the mean raw water oocyst concentration, and (3) knowledge of the variability in concentration (as opposed to the less useful and potentially misleading knowledge associated with variability in concentration estimates) can facilitate detailed risk analyses. If such an approach were to be used in future rounds of monitoring associated with the LT2ESWTR (USEPA, 2006), then the bin classifications would need to be revised because they presently correspond to average counts per unit volume and not to actual concentrations of pathogens in the water. These proposed changes would not address infectivity (they essentially assume that all observed oocysts are infectious), but the present regulation is susceptible to the same problems. For example, a public water system with abundant non-infectious oocysts in its water supply would require more treatment than a public water system that has fewer oocysts that are predominantly infectious. This is also an important problem to address in the future as methods evolve to enumerate only infectious pathogens or to accurately estimate the fraction of pathogens that are infectious in large source water samples).

6.5 QUANTITATIVE MICROBIAL RISK ASSESSMENT: ARE CONCENTRATION DATA BEING HANDLED CORRECTLY?

Quantitative microbial risk assessment (QMRA) is a tool that is used to evaluate the risk to humans as a result of exposure to pathogens. QMRA is a diverse field of research that addresses many types of pathogens, exposure pathways, and applications (Haas *et al.*, 1999). The objective herein is not to provide a comprehensive discussion of QMRA or to apply it to a particular problem, but to describe how measurement errors in microbial enumeration data may affect QMRA results and to evaluate the statistical approaches that are presently in use. The focus of this discussion is upon consumption of drinking water that contains *Cryptosporidium* and *Giardia*, but the discussion will also apply to some extent for other pathogens and other exposure pathways. Discussion of the applications of QMRA in this context and some case studies of its implementation can be found in Regli *et al.* (1991), Gale (1996), Haas *et al.* (1999) and Medema *et al.* (2003).

The general premise of QMRA is to estimate the risk associated with a particular water supply by quantifying the doses to which consumers may be exposed and by quantifying the probability of infection associated with a particular dose. The risk is often reported in terms of infections per 10,000 consumers per year (*e.g.* Regli *et al.*, 1991), and infection is defined as pathogen multiplication within the host. Other measures of risk include morbidity and mortality (Haas *et al.*, 1993), and disability-adjusted life years (Havelaar *et al.*, 2000). The dose, for this particular exposure pathway, is the product of the volume of water consumed and the concentration of the pathogen of interest in the water. Dose-response experiments are conducted to evaluate the probability of infection that is associated with a particular dose (*e.g.* Rendtorff, 1954; Dupont, *et al.*, 1995). The most commonly used dose-response models for this type of application are the exponential and beta-Poisson dose-response models (Haas, 1983). These two models (which have also been discussed in Section 4.2.8 and 5.4.1) both use a Poisson model to relate the pathogen concentration and volume consumed to a number of consumed pathogens. Some researchers have assumed that doses are necessarily integers (*e.g.* Gale *et al.*, 1997), which is incompatible with the exponential and beta-Poisson models (because the models already include a component to address the conversion from concentration and volume to an integer). Once a dose-response model has been estimated from the results of a dose-response experiment (*e.g.* Haas, 1983; Rose *et al.*, 1991a; Regli *et al.*, 1991; Haas *et al.*, 1996; Teunis and Havelaar, 1999) the objective is to quantify the doses to which consumers are typically exposed.

In addition to problems associated with modelling the quantity of unboiled water that is imbibed by a typical consumer, the problem with evaluating doses is that the concentration of pathogens in a safe drinking water supply is too low to measure reliably. QMRA has been used to demonstrate that samples as large as 100,000 L would be needed to demonstrate the safety (in terms of *Cryptosporidium* and *Giardia*) of a treated drinking water supply (Regli *et al.*, 1991, Rose *et al.*, 1995; Haas *et al.*, 1996). Monitoring treated drinking water for potentially pathogenic (oo)cysts using currently available methods has also been criticised because the methods are expensive, inefficient, and yield unreliable data (Allen *et al.*, 2000; Signor and Ashbolt, 2006). Accordingly, most QMRA models have evaluated risks by integrating information about the raw water pathogen concentration and the efficiency of the treatment process. Other issues that must be addressed include infectivity (because enumeration methods generally do not distinguish between infectious and non-infectious pathogens) and analytical recovery (because enumeration methods for (oo)cysts generally do not result in observation of all the (oo)cysts that are present in a sample). Equation 6.12 (Teunis *et al.*, 1997) is commonly used to relate these various factors to dose. In this equation, D is the dose, C is the (observed) concentration of pathogens in the raw water, R is the analytical recovery of the enumeration method, I is the fraction of the pathogens that are infectious, DR is the decimal reduction (called log-reduction throughout this thesis), and V is the volume of treated drinking water imbibed by the consumer. This type of equation represents a sequence of arithmetic conversions (Regli *et al.*, 1991) that essentially calibrate the count per unit volume in the source to the concentration in the treated water.

$$D = C \times 1/R \times I \times 10^{-DR} \times V \tag{6.12}$$

Given a point estimate of each parameter, the dose can be calculated explicitly and so too can the risk; however, many of these parameters vary over time (or among consumers) and there is also uncertainty associated with these parameters (or the variation therein). Accordingly, QMRA has transitioned from point estimates of risk to Monte Carlo analyses that address variability and uncertainty using distributions for each of the parameters in the risk model (*e.g.* Haas *et al.*, 1993; Medema *et al.*, 1995; Teunis *et al.*, 1997; Gale, 1998; Haas *et al.*, 1999; Teunis and Havelaar, 1999; Masago *et al.*, 2002; Medema *et al.*, 2003; Pouillot *et al.*, 2004; Signor and Ashbolt, 2006; Smeets *et al.*, 2007; Jaidi *et al.*, 2009; Cummins *et al.*, 2010).

The Monte Carlo QMRA process generally consists of fitting a distribution to data for each parameter in the model. If data are not available for a particular parameter, typical values are often obtained from the literature or a hypothetical distribution is used. The former is more typical when attempting to quantify the risk associated with an actual water supply, while the latter is more common when evaluating the risk associated with hypothetical scenarios. It is often implicitly assumed in Monte Carlo QMRA that the parameters are all independent. The observed concentration of pathogens (*i.e.* the count divided by the volume) is, however, necessarily dependent upon analytical recovery (Schmidt and Emelko, 2010a; Section 6.4.2). There are also concerns that the log-reduction of a treatment process may be dependent upon the raw water pathogen concentration (*e.g.* Assavasilavasukul *et al.*, 2008). Haas (1999) discussed Monte Carlo modelling approaches that address correlation between parameters. Schmidt and Emelko (2010a) demonstrated that correlation is necessarily case-specific (also demonstrated in Section 6.4.2) and that it is inappropriate to apply an observed level of correlation from one situation to another. Moreover, as demonstrated in Section 6.4.2, most Monte Carlo models have inappropriately regarded analytical recovery as a random variable in the model when it is actually a parameter that relates to the uncertainty of concentration estimates.

Monte Carlo QMRA may use distributions to represent the uncertainty in parameters that are not measured, or to address uncertainty in the hyperparameters that describe variability in a parameter (*e.g.* ρ,λ are hyperparameters for concentration in the models presented in this chapter) using second-order Monte Carlo (*e.g.* Pouillot *et al.*, 2004; Jaidi *et al.*, 2009). Uncertainty in the ‘measured’ concentration and log-reduction values is generally not even raised as a concern that needs to be addressed. Teunis and Havelaar (1999) and Petterson *et al.* (2007) have addressed uncertainty in concentration estimates and in the parameters of the gamma distribution that describes temporal concentration variability by using Bayes’ theorem (as discussed, and compared to the models presented herein, in Section 6.1.2).

Schmidt and Emelko (2010a) argue that “appropriate QMRA model development requires accurate conceptual and quantitative distinction between ‘variability’ and ‘uncertainty’”. The distinction is discussed herein in Section 3.1.1. Haas *et al.* (1999) described many different types of variability and uncertainty that must be addressed in QMRA. In the case of raw water pathogen concentrations, for example, Schmidt and Emelko (2010a) argued that the following types of variability and uncertainty must be addressed to evaluate temporal variability among pathogen concentration data correctly:

"(1) uncertainty in the concentration estimates, (2) temporal variability of the pathogen concentration, and (3) uncertainty in the choice of temporal concentration variability model and its estimated parameters". All three of these were addressed in Schmidt and Emelko (2010a) and they are addressed herein in the discussion of the appropriateness of the gamma distribution to model temporal concentration variability (Section 6.1.4) and the Gibbs sampling approach presented in Section 6.3.1.

The objectives of this section are to summarize commonly used approaches to model temporal concentration variability and to address analytical recovery (Sections 6.5.1 and 6.5.2 respectively), to contrast these with the Bayesian approach presented herein (Section 6.5.3), and to discuss other elements of the QMRA process in which measurement errors associated with microbial enumeration must be considered (Section 6.5.4).

6.5.1 Fitting Distributions to Temporally Distributed Concentration Estimates

Many Monte Carlo QMRA models have fitted distributions describing temporal variability to sets of concentration estimates (*e.g.* Gale, 1998; Jaidi *et al.*, 2009). Fitting distributions to sets of enumeration-based concentration estimates to describe temporal variability implicitly assumes that the concentration estimates are exact; the measurement errors in the enumeration process and the associated uncertainty in each concentration estimate are disregarded. Other Monte Carlo QMRA models (*e.g.* Medema *et al.*, 1995; Teunis *et al.*, 1997; Medema *et al.*, 2003) have fitted distributions to the counts of observed microorganisms. Medema *et al.* (1995), Teunis *et al.* (1997) and Medema *et al.* (2003) have not explained how the counts were converted to concentrations in the Monte Carlo risk analyses. Fitting distributions to counts assumes that the counts vary due to temporal variability in concentration alone and not due to other measurement errors (*e.g.* random sampling error and analytical error). Schmidt and Emelko (2010a) demonstrated that dividing concentration estimates by analytical recovery values removes the bias due to analytical error, but it does not remove the uncertainty. In fact, dividing concentration estimates by random recovery values from a non-constant analytical recovery distribution inflates errors rather than addressing uncertainty (Section 6.4.2).

A final problem with fitted distributions that do not address measurement errors in the enumeration method is that "routine analytical and experimental issues such as non-detect samples, variable sample volumes, variable recovery information (*e.g.* different enumeration methods), and replication pose significant difficulties" (Schmidt and Emelko, 2010a). Fitting a distribution that describes temporal variability to sets of counts or concentration estimates assumes that the data are independent and identically distributed. This assumption is untrue unless all data are obtained using equal sample volumes and the same enumeration method (*i.e.* with the same variability in analytical recovery). Moreover, it is proven in Section 8.2.3 that counts and concentration estimates associated with different source concentrations do not have equal error (*e.g.* higher concentrations yield more precise concentration estimates when the sample volume and analytical recovery are the same). Non-detect samples are problematic because log-transforming concentration estimates, calculating geometric means, or fitting distributions to concentration estimates that do not allow values of zero are not possible unless zeros are manipulated into positive values. Numerous schemes have been used to manipulate non-detect results (*e.g.* changing the count to a value of 0.5 or 1, or using a distribution of positive values for non-detects). It is typically assumed that the concentration is actually a positive value that the analyst randomly failed to detect (*e.g.* due to poor sensitivity of the enumeration method). All counts (zero or otherwise) correspond to uncertain concentration estimates, and this uncertainty has generally been disregarded.

6.5.2 Incorporating Analytical Recovery into QMRA Models

It has been demonstrated (*e.g.* Sections 5.4.4, 6.2, and 8.2.1) that concentration estimates obtained by enumeration are biased unless they are calibrated to actual concentrations using information about the analytical recovery of the method. Accordingly, many QMRA models have sought to remove this bias by dividing the count per unit volume by analytical recovery (*e.g.* Equation 6.12). Nonetheless, analytical recovery has often been ignored and Schmidt and Emelko (2010a) have attributed this to three reasons: "(1) losses in the method have not been considered or were assumed to be negligible (*e.g.* plating methods, most probable number methods, molecular methods), (2) recovery was acknowledged as an important factor but appropriate recovery data were not available (Rose *et al.*, 1991a), or (3) under-estimation of microbial concentrations due to incomplete recovery was assumed to counteract over-estimation of the abundance of infectious microorganisms (Regli *et al.*, 1991; USEPA, 2006; Smeets *et al.*, 2007)". Monte Carlo analyses that have addressed analytical recovery have generally regarded it as an independent random variable with respect to microorganism counts or concentration estimates. This section shows that analytical recovery and infectivity are unlikely to counter-balance each other when evaluating temporal variability of infectious pathogen concentrations, that analytical recovery and microorganism counts are not independent even if they are uncorrelated, and that incorporating analytical recovery into Monte Carlo QMRA as an independent (or even as a dependent) random variable artificially inflates the variability of concentration.

Analytical recovery and infectivity are unlikely to counter-balance each other

Infectivity is an important issue in risk analysis because pathogens that are enumerated using a method that does not distinguish between infectious and non-infectious microorganisms may count some microorganisms that pose no threat to consumers. For example, Method 1623 (USEPA, 2005b) can be used to determine how many of the total observed (oo)cysts have a viable type morphology, but cannot be used to determine whether or not the observed (oo)cysts are of a species/genotype that is pathogenic to humans, or if the observed (oo)cysts have somehow become inactivated. The rationale behind the assertion that the errors associated with analytical recovery and infectivity counteract is that analytical recovery below 100% causes the count per unit volume to be an under-estimate of concentration while infectivity below 100% causes the pathogen concentration to be an over-estimate of the concentration of infectious pathogens. As proven below, the assertion that the two errors counter-balance (in terms of the mean and variance of concentration estimates without regard to the shape of the distribution) is only true under very strict conditions. In general, it is inappropriate to assume that the two errors counter-balance.

Mathematical expectation and variance decomposition (Section 3.3.2) are used herein to investigate the assumption that these errors counter-balance each other. It is assumed that the concentration of infectious microorganisms (c_I) is related to the concentration of enumerable microorganisms (c) by the equation $c_I = i \times c$, in which i is "the fraction of enumerated microorganisms that are viable and of epidemiologically relevant species/genotype" (Schmidt and Emelko, 2010a). It is also assumed that random sampling and analytical error are represented by a Poisson distribution with mean cVp (in which V is the sample volume and p is the analytical recovery) and that infectivity and analytical recovery are independent. At a minimum, the assertion that the two errors counter-balance essentially assumes that the count (x) per unit volume is on average equal to the concentration of infectious microorganisms and that the variance of the counts per unit volume is equal to the variance of the concentration of infectious pathogens. Equations 6.13 and 6.14 show the conditions under which each assumption, respectively, is met. In these equations, μ_p and σ_p^2 represent the mean and variance of analytical recovery (among samples), μ_I and σ_I^2 represent the mean and variance of infectivity (among sampling events), and μ_c and σ_c^2 represent the mean and variance of the concentration including both infectious and non-infectious microorganisms (among sampling events). Derivation of these equations is provided in Appendix E.

$$E\left[\frac{x}{V}\right] - E[c_I] = \mu_c \cdot \left(\mu_p - \mu_I\right) = 0 \tag{6.13}$$

$$Var\left[\frac{x}{V}\right] - Var[c_I] = \mu_c \frac{\mu_p}{V} + \left(\mu_c^2 + \sigma_c^2\right)\sigma_p^2 + \mu_p^2\sigma_p^2 - \left(\mu_c^2 + \sigma_c^2\right)\sigma_I^2 = 0 \tag{6.14}$$

Equation 6.13 shows that the count per unit volume is only an unbiased estimate of the concentration of infectious pathogens if $\mu_p = \mu_I$. The variability of counts per unit volume is equal to the temporal variability of the concentration of infectious pathogens if Equation 6.14 is true. If both equations are not satisfied, then ignoring analytical recovery and infectivity will result in biased concentration estimates, a biased estimate of the temporal variability of the infectious pathogen concentration, or both. It is better to address both of these errors in the risk analysis because Equations 6.13 and 6.14 are difficult to prove (they depend upon means and variances that cannot be estimated precisely without considerable amounts of data and appropriate statistical analysis) and because the distributions of the two errors may have different shape even if Equations 6.13 and 6.14 are satisfied. The models and Bayesian analyses presented in Sections 6.1 and 6.2 respectively address analytical recovery but not infectivity. These modelling approaches only assess the concentration of infectious pathogens if it is assumed that 100% of the observed pathogens are infectious. Infectivity can be addressed separately in a Monte Carlo QMRA if it is assumed that the fraction of microorganisms that are infectious is independent of the estimated concentrations (and the analytical recovery of the enumeration method). "Quantification of the infectivity of observed microorganisms (and the error therein) has not yet been fully addressed and requires continued research and consideration by laboratories, modellers, and regulators" (Schmidt and Emelko, 2010a). Methods to evaluate the infectivity and genotype of observed *Cryptosporidium* oocysts (*e.g.* Aboytes *et al.*, 2004; Di Giovanni *et al.*, 2009) have been proposed, but are not in wide-spread use.

Microorganism counts and analytical recovery are not independent random variables

> "Many approaches to integrate analytical recovery information into the analysis of microorganism count data have (often implicitly) treated microorganism counts and analytical recovery as statistically independent values (*e.g.* Medema *et al.*, 1995; Teunis *et al.*, 1997; Teunis and Havelaar, 2002; Medema *et al.*, 2003; Pouillot *et al.*, 2004; Signor and Ashbolt, 2006; Cummins *et al.*, 2010) or have used built-in features in commercially available Monte Carlo-based software to apply user-specified levels of correlation (Jaidi *et al.*, 2009; Prévost and Barbeau, personal communication). The number of microorganisms observed in a sample, however, is necessarily dependent upon analytical recovery because fewer microorganisms will be observed (on average) when recovery is lower." – Schmidt and Emelko (2010a)

The dependence between counts and analytical recovery was evaluated by Schmidt and Emelko (2010a) using Monte Carlo simulations based upon the beta-Poisson temporal concentration variability model (Section 6.1.1) with internal seed recovery data simulated also. The simulations assumed sample volumes of 100 L or 5 L and internal seed doses of 100 microorganisms, and they used a variety of values for the temporal concentration variability distribution parameters ρ,λ and the non-constant analytical recovery distribution parameters α,β. The simulations used three alternative temporal concentration variability distributions that had a mean of 10 microorganisms/L and different standard deviations (10, 5, and 1 microorganisms/L respectively). The parameters (ρ,λ) of these distributions were (1,10), (4,2.5), and (100,0.1) respectively. Two alternative recovery distributions were used: the first (with $a = 3$, $b = 7$) has a mean of 30% and a standard deviation of 13.8%, and the second (with $a = 70$, $b = 30$) has a mean of 70% and a standard deviation of 4.6%. A total of 8 simulations were conducted (with 24 data each): 6 simulations representing the combinations of the temporal concentration variability distributions and non-constant analytical recovery distributions with sample volumes of 100 L, and 2 simulations using 5 L sample volumes and $(\rho,\lambda) = (100,0.1)$. To illustrate correlation between counts and analytical recovery, scatter plots of the counts of indigenous microorganisms (x) and the counts of seeded microorganisms (x^*) are shown in Figure 6.5. The counts of indigenous microorganisms are plotted against counts of seeded microorganisms rather than the analytical recovery values because the recovery associated with each sample is unknown in practice and the fraction of seeded particles that are observed (x^*/n^*) is presumed to be a somewhat precise estimate of sample-specific analytical recovery. These results show that the counts of indigenous and seeded microorganisms become increasingly correlated (*i.e.* the trend of the plotted data has a steeper slope) as the variability in concentration decreases, as the variability in analytical recovery increases, or as the sample volume increases. Correlation "reflects the impacts of variability in recovery upon the indigenous microorganism count data" (Schmidt and Emelko, 2010a) and can easily be confounded by other sources of variability (*e.g.* variability in concentration or increased random sampling error associated with smaller volumes).

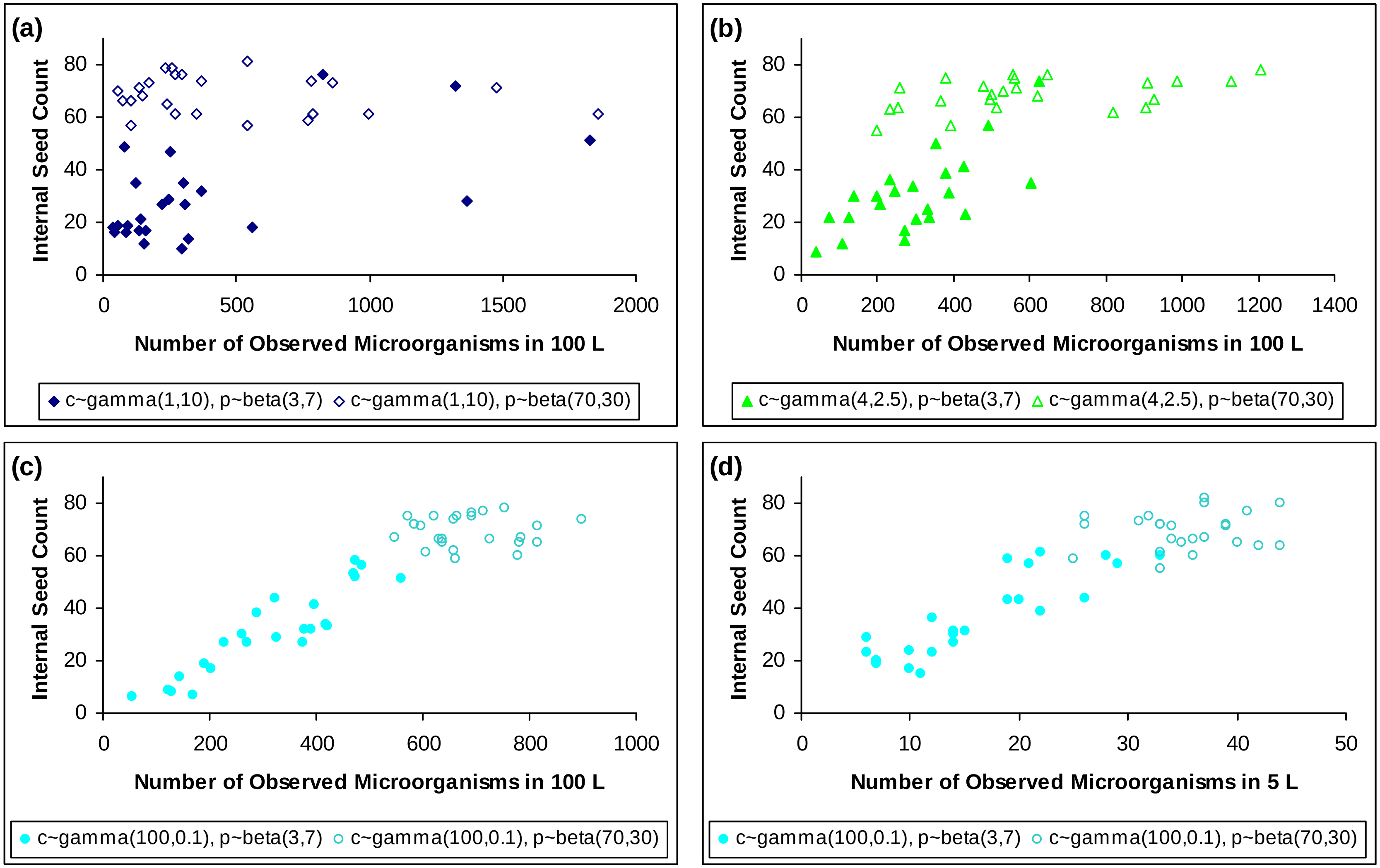


Figure 6.5: Demonstration of correlation between counts and analytical recovery (Schmidt and Emelko, 2010a)

Data simulated using the beta-Poisson temporal concentration variability model together with internal seed recovery data. Panels represent (a) 100 L samples with high temporal concentration variability, (b) 100 L samples with moderate temporal concentration variability, (c) 100 L samples with low temporal concentration variability, and (d) 5 L samples with low temporal concentration variability.

Petterson *et al.* (2007) had a dataset of temporally distributed enumeration data that were each paired with internal seed data. Scatter plots of the data for *Cryptosporidium* oocysts and *Giardia* cysts that were similar to the plots in Figure 6.5 showed that the internal seed recovery estimates were uncorrelated with the counts of indigenous microorganisms.

> "A relationship between counts and recovery should be intuitively expected; given a constant underlying source water concentration, as the recovery increases, so the number of organisms counted would also increase. The fact that this relationship was not observed in these datasets suggests that the influence of variation in recovery was small relative to the variation in underlying [oo]cyst concentration." – Petterson *et al.* (2007)

While correlation demonstrates dependence, the absence of correlation is insufficient evidence that two variables are statistically independent.

> "Statistical dependence means that the distribution of possible values of one random variable is influenced by the value of another. Microorganism counts and analytical recovery are dependent because the number of observed microorganisms depends upon the probability that each microorganism in a sample will be observed (analytical recovery). Correlation between microorganism counts and analytical recovery is merely an indicator of this omnipresent dependence." – Schmidt and Emelko (2010a)

It is concluded in Schmidt and Emelko (2010a) that "statistical analyses that treat microorganism counts and analytical recovery as independent will over-predict the variability in concentration by pairing extreme numbers of observed microorganisms with extreme values of analytical recovery (*e.g.* unrealistically high concentration estimates may be obtained by pairing relatively high counts with relatively low recovery values)" and that "observed levels of correlation are case-specific and should not be applied from one data set to another (because changing any factor that affects variability in the data will change the correlation)". The Bayesian approach (*e.g.* Section 6.3) to evaluate temporal concentration variability and the uncertainty therein was recommended as a strategy to appropriately address measurement errors in applications such as QMRA.

Analytical recovery is not a random variable in QMRA

Monte Carlo QMRA analyses that have addressed analytical recovery have typically regarded it as a random variable (usually an independent random variable) and have integrated its variability into the model using a distribution of recovery values. Analytical recovery, however, is not a random variable in this context. If the concentration of infectious pathogens (C_I) could be precisely enumerated in the raw water, then the dose model (simplified from Equation 6.12) would be Equation 6.15. In this model, it is assumed that C_I and the log-reduction (DR) are temporally distributed random variables (which may potentially be dependent), and that the consumed volume (V) is also a random variable among consumers.

$$D = C_I \times 10^{-DR} \times V \qquad (6.15)$$

The problem with this model is that enumeration methods may also count non-infectious pathogens and that the enumeration methods may have analytical errors (*i.e.* analytical recovery may not be consistently 100%). If the enumeration method also counts non-infectious pathogens, then C_I cannot be estimated directly and Equation 6.16 may be used to evaluate exposure. The fraction of the pathogens in the water supply that are infectious (I) may be dependent upon the concentration of the enumerable pathogens (C), which includes both infectious and non-infectious pathogens.

$$D = C \times I \times 10^{-DR} \times V \qquad (6.16)$$

Analytical recovery is not a random variable in this model; rather, it is a factor that affects the uncertainty in (or imprecision of) the individual concentration estimates. This is graphically illustrated in Figure 6.6. In the Bayesian approach to evaluate temporal concentration variability with appropriate consideration of measurement errors, analytical recovery is regarded as an unknown parameter (with a prior that may be described using a non-constant analytical recovery distribution) Regarding analytical recovery as a random variable (whether independent or dependent), rather than as an unknown parameter in the investigation of measurement errors, will artificially inflate the variability of concentration estimates rather than addressing the fraction of the variability among the concentration estimates that is due to measurement errors. Accordingly, Monte Carlo QMRA analyses that regard analytical recovery as a random variable will not only over-estimate temporal concentration variability due to the imprecision of the concentration estimates, but will further inflate the variability by falsely regarding analytical recovery as a random variable.

To graphically demonstrate this concept, the analyses in Section 6.2 were repeated and a series of concentration estimates was generated by dividing the simulated count by a random analytical recovery value. The results (Figures 6.7 and 6.8) show that dividing the count per unit volume by a random analytical recovery value (as opposed to the mean analytical recovery) will substantially increase the variability of the concentration estimates in most cases. In general, (1) concentration estimates that are not adjusted for analytical recovery are biased, (2) concentration estimates that are adjusted for mean analytical recovery are unbiased, but more variable than the concentration itself, and (3) concentration estimates that are divided by random analytical recovery values are unbiased, but are even more variable than the concentration estimates that accounted for mean analytical recovery. Table 6.2 presents the 99th percentile values of the various concentration estimates and shows that d counts per unit volume by random recovery values can result in unrealistic high concentration values. These high concentration values have the greatest impact upon overall risk (when temporal variability in concentration is addressed) and will therefore result in over-predicted risks. Similar inflation of variability occurs using the approach proposed by Pouillot *et al.* (2004). Figure 6.9 shows a scenario in which 10,000 simulated samples (each containing 25 microorganisms) are enumerated by a method with beta-distributed analytical recovery (a = 2.65, b = 3.64) and each count is subsequently divided by a random analytical recovery value from the same distribution. It is clear that the method proposed by Pouillot *et al.* (2004) artificially inflated variability in count data rather than appropriately evaluating the uncertainty in the concentration estimates.

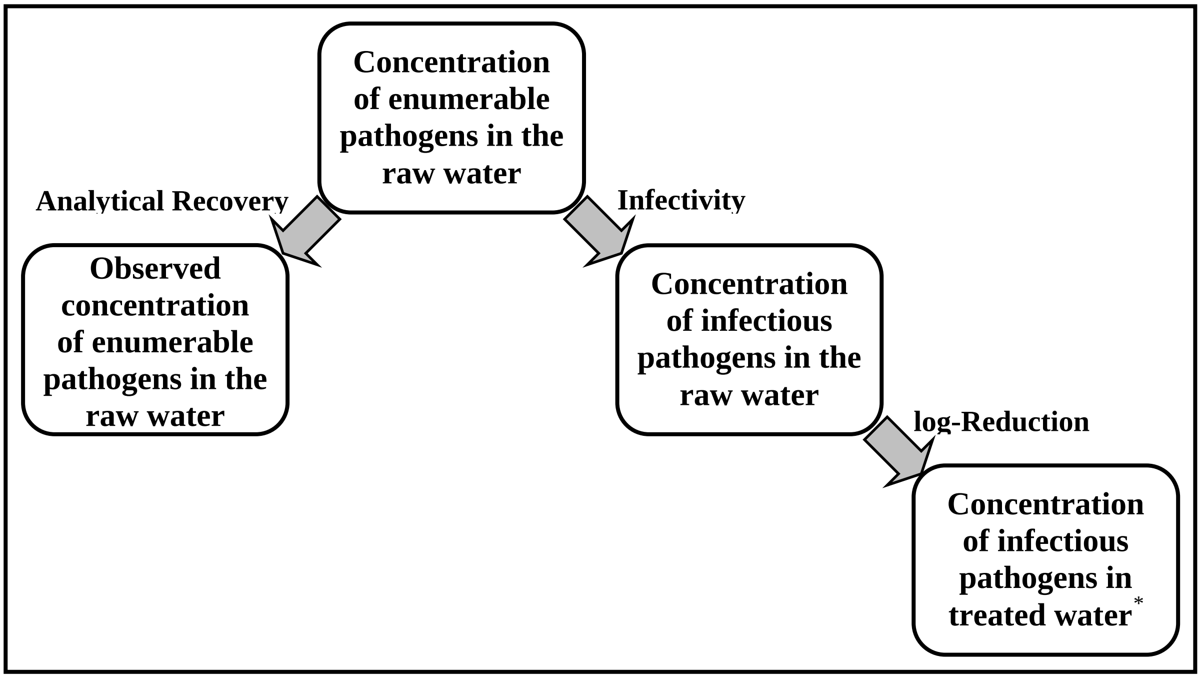

Concentration of enumerable pathogens in the raw water
Analytical Recovery
Infectivity
Observed concentration of enumerable pathogens in the raw water
Concentration of infectious pathogens in the raw water
log-Reduction
Concentration of infectious pathogens in treated water*

Figure 6.6: Errors in evaluating treated water infectious pathogen concentrations

When estimating the variability of treated water infectious pathogen concentrations, analytical error due to imperfect analytical recovery results in uncertainty in the raw water concentration of enumerable pathogens, while variability in infectivity and log-reduction are errors that result in variability in the treated water infectious pathogen concentration.

*Some researchers such as LeChevallier (2009) have addressed additional variability due to the distribution system (*e.g.* variability in tap water concentrations in time and throughout the system due to mixing, infiltration, *etc.*), but such analyses are usually hypothetical and not based upon actual monitoring data.

Table 6.2: 99th percentile values of various concentration estimation procedures

Figure	99th Percentile Concentration Value (microorganisms/L)			
	True Concentration	'No Recovery' Conc. Estimate	'Mean Recovery' Conc. Estimate	'Random Recovery' Conc. Estimate
6.7a	46.582	17.400	58.000	98.810
6.7b	44.640	14.200	47.333	49.661
6.7c	46.391	34.400	49.143	54.529
6.7d	46.655	17.000	56.667	109.999
6.8a	12.514	7.100	23.667	71.446
6.8b	12.436	5.000	16.667	18.991
6.8c	12.497	11.100	15.857	23.090
6.8d	12.460	6.940	23.133	65.171

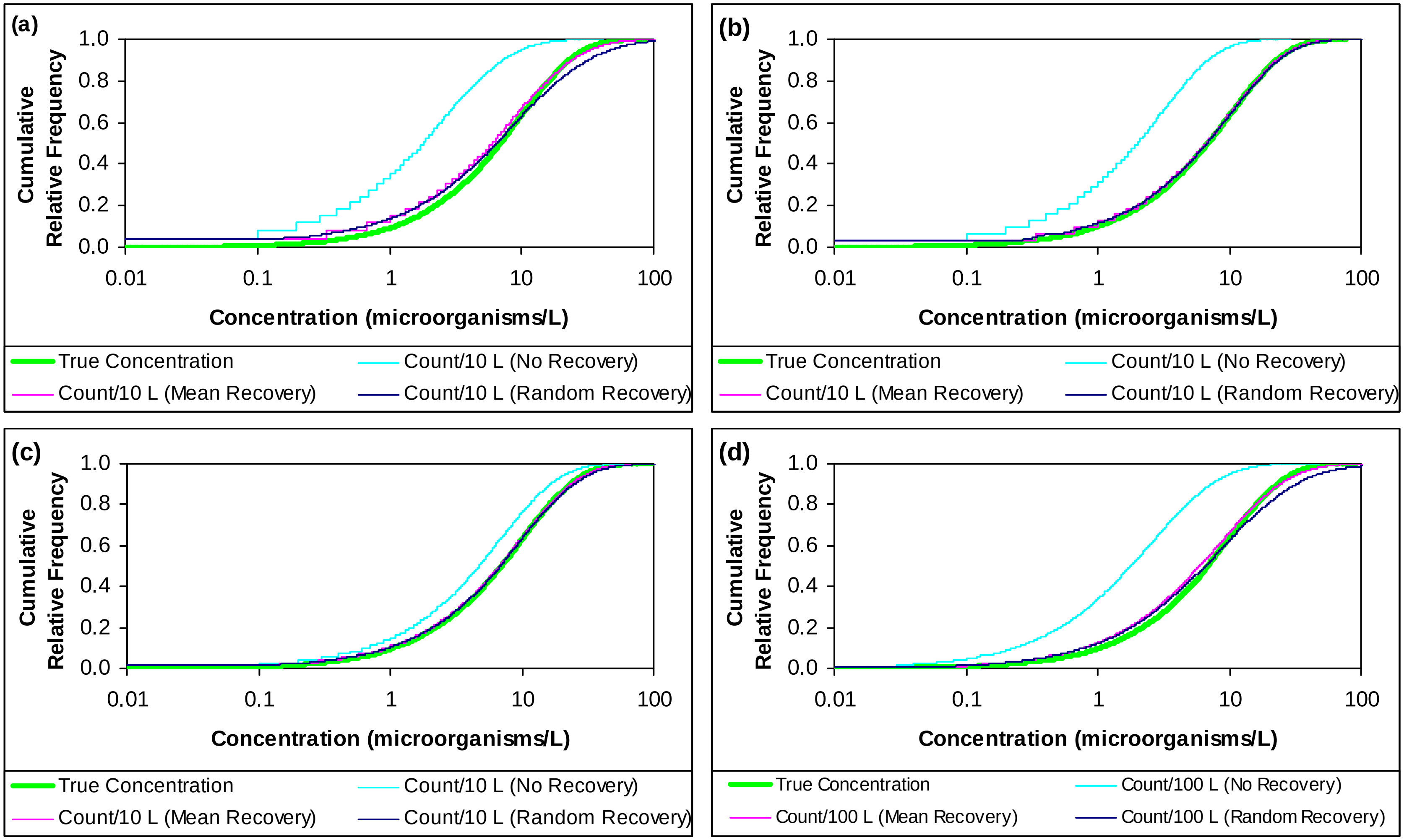


Figure 6.7: Effects of regarding analytical recovery as a random variable when temporal concentration variability is high

These analyses are based upon 10,000 simulations using $\rho = 1$, $\lambda = 10$ ($\mu_c = 10$, $\sigma_c = 10$ in microorganisms/L). The panels represent (a) a base scenario against which other panels are compared [$a = 3$, $b = 7$, $V = 10$], (b) a scenario with less variable analytical recovery [$a = 30$, $b = 70$, $V = 10$], (c) a scenario with higher mean recovery [$a = 7$, $b = 3$, $V = 10$], and (d) a scenario with higher sample volume [$a = 3$, $b = 7$, $V = 100$]. ‘No Recovery’, ‘Mean Recovery’ and ‘Random Recovery’ describe what type of adjustment for recovery is used.

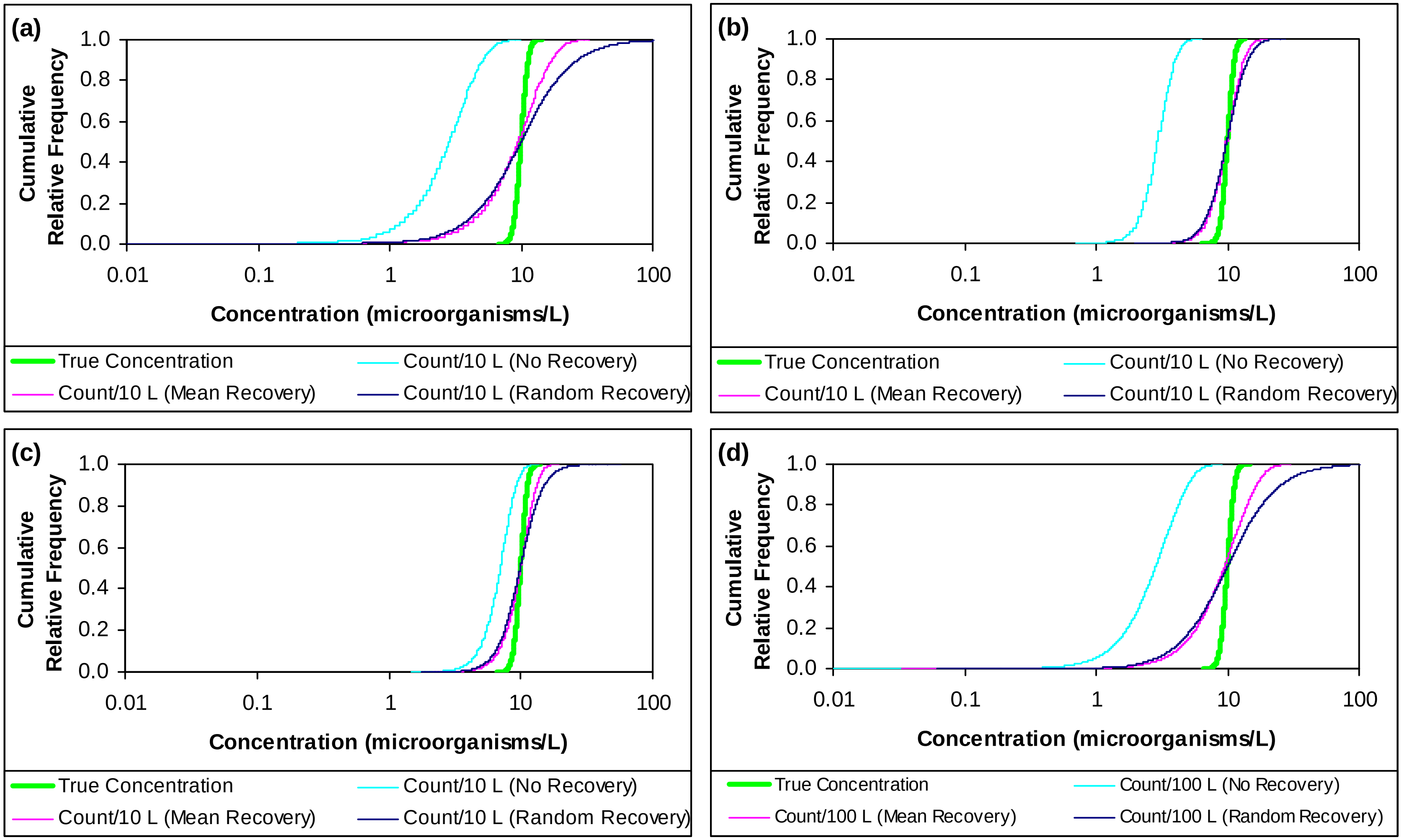


Figure 6.8: Effects of regarding analytical recovery as a random variable when temporal concentration variability is low

These analyses are based upon 10,000 simulations using ρ = 10, λ = 1 (μ_c = 10, σ_c = 1 in microorganisms/L). The panels represent (a) a base scenario against which other panels are compared [a = 3, b = 7, V = 10], (b) a scenario with less variable analytical recovery [a = 30, b = 70, V = 10], (c) a scenario with higher mean recovery [a = 7, b = 3, V = 10], and (d) a scenario with higher sample volume [a = 3, b = 7, V = 100]. ‘No Recovery’, ‘Mean Recovery’ and ‘Random Recovery’ describe what type of adjustment for recovery is used.

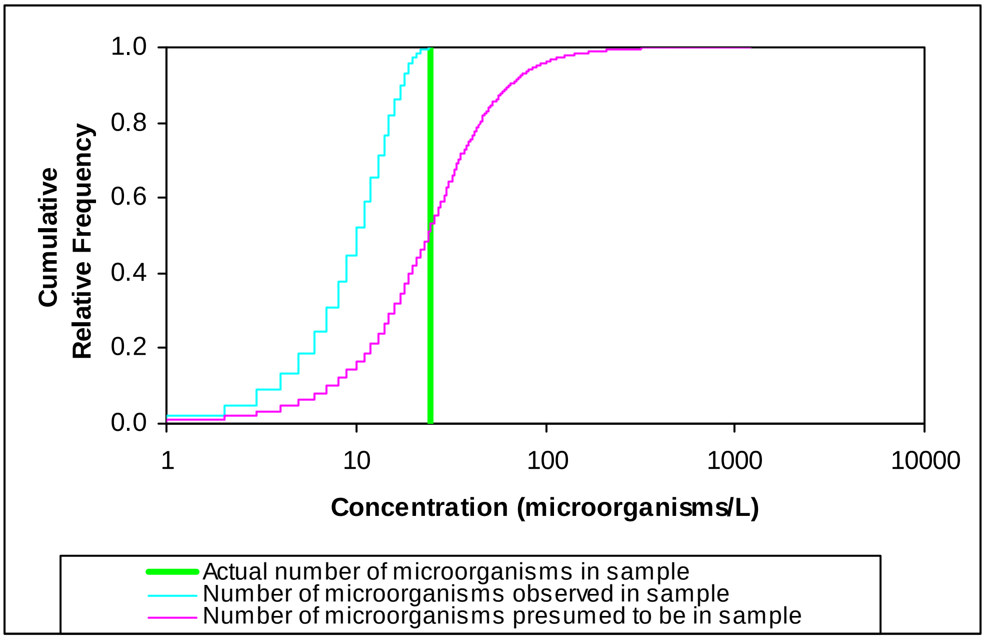


Figure 6.9: Effects of adjusting microorganism counts by random recovery values

Results based upon 10,000 simulations with 25 particles per sample and beta-distributed recovery with *a* = 2.65, *b* = 3.64. The final series represents counts of observed microorganisms that have been adjusted for analytical recovery using the negative binomial approach proposed by Pouillot *et al.* (2004).

6.5.3 Comparison of Conventional Monte Carlo Data Analyses with a Bayesian Approach

Monte Carlo QMRA models that fit distributions to concentration estimates or counts ignore the effects of measurement errors upon the distribution of the data and often give special treatment to non-detect samples (Section 6.5.1). The simulations shown in Section 6.2 demonstrate that Monte Carlo QMRA models that ignore analytical recovery (when it is below 100% on average) will under-estimate concentrations. Finally, it was demonstrated in Section 6.5.2 that it is not acceptable to regard analytical recovery as a random variable in Monte Carlo QMRA models: although it removes bias in the mean concentration by calibrating concentration estimates, it adds bias to the variance. Temporally distributed enumeration data can only be appropriately analyzed using statistical procedures that properly address all of the measurement errors in the data. The probabilistic models presented in Section 6.1.1 address the measurement errors in enumeration data, and the Bayesian analyses presented in Section 6.3.1 provide a framework to evaluate the temporal variability in concentration (and the uncertainty therein) with appropriate consideration of measurement errors.

This section compares the Bayesian approach to evaluate temporal concentration variability to alternative approaches that are comparable to conventional Monte Carlo QMRA procedures. Specifically, the simulated data summarized in Appendix D.4 (which were analyzed using Bayesian methods in Section 6.3.2) are analyzed using two Bayesian approaches and two bootstrapping approaches. The Bayesian approaches are based upon the beta-Poisson temporal concentration variability model (which ignores the internal seed recovery data in the dataset and uses a beta distribution with parameters $a = 2$, $b = 3$ to address non-constant analytical recovery) and the internal seed temporal concentration variability model (which uses the internal seed recovery data and uniform priors for each sample's probability of recovery). In each case, the posterior predictive distribution of concentration is used to represent the available knowledge about variability in analytical recovery. The posterior predictive distributions are computed as illustrated in Figure 6.3 using the Markov chains of parameter values that are plotted in Figures 6.4a and 6.4b and a burn-in of 1000 iterations.

> "Two bootstrapping approaches ('with zeros' and 'no zeros') were also used to compare the Gibbs sampling results with conventional approaches that assume that microorganism counts and analytical recovery are independent. For each boot-strapping iteration, one of the counts (x_i) was drawn at random and then divided by its processed sample volume and a random value of analytical recovery (drawn from the beta distribution) to yield a concentration estimate. In the 'no zeros' approach, non-detect samples were assigned a count of 0.5 as an example of analyses in which non-detect samples are manipulated into non-zero values." – Schmidt and Emelko (2010a)

Figure 6.10 shows the results of these four analyses compared to the temporal concentration variability distribution from which the data were simulated. It is clearly demonstrated in Figure 6.10a that bootstrapping or similar fitting-based Monte Carlo approaches are complicated by abundant non-detect samples. The 'with zeros' bootstrapping approach does not properly address temporal concentration variability because it ascribes a concentration value of zero to all non-detect samples when the concentration is presumably some low positive value that randomly gave rise to a non-detect sample (because the sample volume was too small or the enumeration method had low analytical recovery). In contrast, the 'no zeros' bootstrapping approach is biased because it manipulates non-detects into positive values that are often much higher than the actual concentration. The two Bayesian posterior predictive distributions illustrated in Figure 6.10b provided a much better description of temporal concentration variability because they appropriately addressed the measurement errors in the enumeration process that cause concentration estimates to be uncertain. The slight discrepancy between the two posterior predictive concentration distributions and the actual concentration distribution is to be expected when temporal concentration variability is evaluated using a small dataset. "In general, the accuracy of posterior predictive distributions obtained from the proposed model (assuming that the model is correct and that the chosen priors are not too restrictive), depends on the number and quality of available data (*i.e.* very inaccurate data contain very little information)" (Schmidt and Emelko, 2010a).

Figure 6.10c and Table 6.3 show various percentiles from each of the 5 distributions. In QMRA, higher raw water concentrations are likely to lead to higher ingested doses, and higher doses are more likely to lead to infection. Accordingly, the higher percentiles of the temporal concentration variability distribution are of interest, and modelling approaches that systematically inflate the

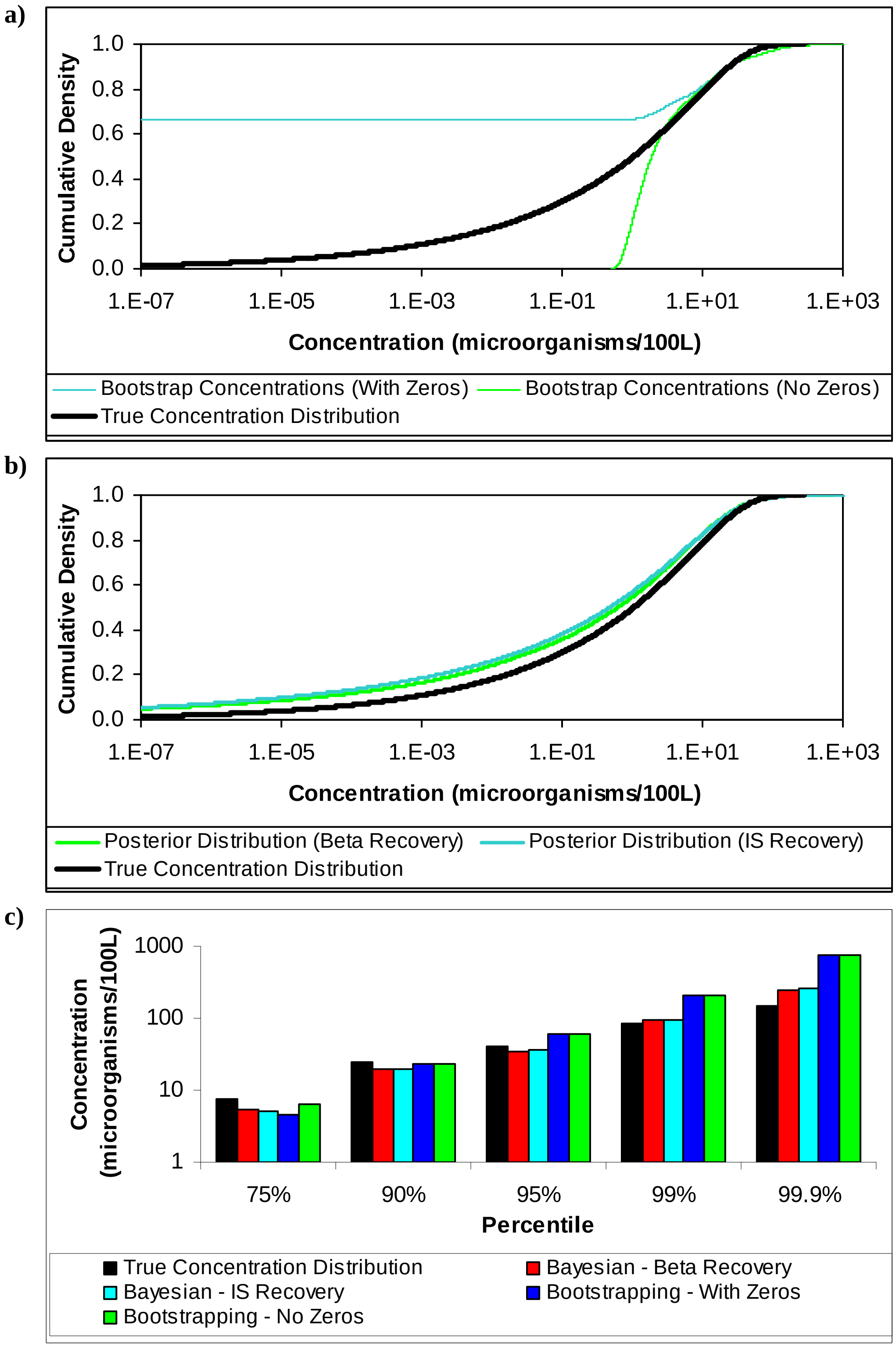


Figure 6.10: Description of temporal concentration variability using several methods
(modified from Schmidt and Emelko, 2010a)

Panels represent (a) comparison of the two bootstrapping approaches with the true concentration distribution from which the data were simulated, (b) comparison of the two Bayesian approaches with the true concentration distribution, and (c) comparison of percentiles from the various distributions.

Table 6.3: Concentration percentiles obtained using various data analysis approaches
(Schmidt and Emelko, 2010b)

Model	75%	90%	95%	99%	99.9%
True Concentration Distribution	7.61	23.9	39.8	82.7	151
Bayesian – Beta Recovery	5.43	19.4	34.4	92.5	253
Bayesian – IS Recovery	5.21	19.6	35.9	95.4	256
Bootstrapping – With Zeros	4.57	22.7	60.3	202	744
Bootstrapping – No Zeros	6.42	23.1	60.6	202	744

probabilities of high concentrations compared to what occurs in reality will over-predict risk. These plotted percentiles show that the two posterior predictive distributions are in relatively good agreement with the actual temporal concentration variability distribution. In contrast, the 95th, 99th, and 99.9th percentiles associated with the two bootstrapping approaches are clearly much greater than those associated with the actual temporal concentration variability distribution. These results further demonstrate that Monte Carlo QMRA procedures that ignore measurement errors in the enumeration process and that regard analytical recovery as a random variable in the model (when it is actually an unknown parameter) are inappropriate. Further research of specific case studies (which address all components of QMRA and not just temporal concentration variability) is required to determine the overall impacts of inappropriate modelling practices upon computed risks.

6.5.4 Measurement Errors in Other Components of QMRA Models

This section describes other components of QMRA models in which the measurement errors discussed in this thesis must be considered. In particular, these are infectivity, decimal (or log-) reduction, and the dose-response model.

Infectivity

The infectivity of pathogens in a water supply (whether detected by an enumeration method or not) must be addressed because only those that are viable and of infectious species/genotypes pose a health risk to consumers. Unless an enumeration method only detects infectious pathogens, the infectivity (the proportion of all enumerable pathogens during a particular sampling event that are infectious) must be addressed to properly evaluate the concentration of epidemiologically important pathogens. Infectivity may be evaluated by determining the fraction of observed pathogens that are infectious. Like the fraction of observed particles with respect to the number actually present, the fraction of observed pathogens that are infectious is only an estimate of infectivity in the source (which may vary among sampling events). Further research will be required to develop methods that evaluate infectivity, to determine whether the information such methods yield warrants routine implementation of more complicated enumeration methodology, and to develop statistical tools that incorporate information about infectivity into the analysis of temporal concentration variability.

Decimal reduction

The decimal reduction (or log-reduction) of a treatment process is difficult to evaluate. Full-scale decimal reduction data for pathogens are generally unavailable because it is desirable for the pathogen concentration to be low in the raw water and practically immeasurable in the treated water. Decimal reduction of pathogen concentrations may be estimated using pilot-scale seeding experiments (*e.g.* Huck *et al.*, 2001; Emelko *et al.*, 2004; Assavasilavasukul *et al.*, 2008), which may include the pathogen or surrogates (*e.g.* polystyrene microspheres, *Bacillus subtilis* spores, and formalin-inactivate oocysts may be considered as transport surrogates for indigenous oocysts). Alternatively, other more plentiful microorganisms can be quantified at full-scale that are believed to be surrogates for pathogen transport or disinfection (*e.g.* Rice *et al.*, 1996). Concerns about the representativeness of decimal reduction data must certainly be addressed in QMRA.

Apart from concerns about the representativeness of available decimal reduction data, measurement errors in the data must also be addressed. As described in Section 5.4, decimal reduction cannot be measured precisely because it is a function of two imprecise concentration estimates. Accordingly, measurement errors must be addressed in experiments and statistical models that are used to quantify decimal reduction. Analyses must address temporal decimal reduction variability (and the uncertainty therein) as well as random sampling and analytical errors and non-constant analytical recovery associated with the enumeration data upon which decimal reduction estimates are based. It was also demonstrated in Section 5.4.4 that log-reduction estimates can be biased if the mean analytical recovery of the enumeration method varies between the treated and untreated waters. Further research is required to appropriately evaluate temporal decimal reduction variability (whether at bench-, pilot-, or full-scale and whether using pathogens or surrogates) with consideration of the measurement errors in the enumeration data. Inappropriate decimal reduction models may substantially over- or under-predict risks to consumers.

Dose-response experiments

Some of the types of measurement errors addressed herein are also important in dose-response experiments. Seeding error is important because the model must address whether the subjects in the experiment were given samples containing precisely known numbers of pathogens or samples drawn from a stock of known concentration (the exponential and beta-Poisson dose-response models assume the latter). Measurement errors in stock concentrations should also be addressed. Finally, the number of subjects (to whom a specific dose was given) in whom infection is detected is a random variable that depends upon the probability of infection associated with that dose and the number of subjects given that dose, so the fraction of subjects who become infected is only an imprecise estimate of the probability of infection. The latter concern is separate from concerns about the representativeness of the subjects with respect to the whole population. Messner *et al.* (2001) have proposed Bayesian approaches to address uncertainty in the parameters of dose-response models, and this uncertainty should also be addressed in Monte Carlo QMRA.

6.6 PRACTICAL IMPLICATIONS

This section takes the concept that the count per unit volume is an imprecise estimate of concentration that must be calibrated to actual concentrations using information about analytical recovery and builds upon it by addressing the situation in which concentration varies randomly over time. The concepts presented herein can also be used to address random spatial variability or a blend of random spatial and temporal variability. The presented probabilistic model is more generalized than other comparable models (because it enables replication within sampling events and sample-specific recovery information), and such models should be used when simulating temporally distributed enumeration data or developing statistical approaches to analyze collected enumeration data.

Monte Carlo simulation of temporally distributed enumeration data confirms that the count per unit volume is a biased concentration estimate unless analytical recovery is consistently 100%, and that the count per unit volume divided by the mean analytical recovery (or a sample-specific estimate of recovery) is an unbiased concentration estimate. Even unbiased concentration estimates, however, are imprecise, and the variability due to measurement errors will cause these concentration estimates to be more variable than the actual concentration. In general, more precise concentration estimates can be obtained by increasing mean analytical recovery, decreasing the variability of analytical recovery, increasing sample volumes, or (if analytical recovery is highly variable) by obtaining internal seed recovery estimates. In some cases, however, developing and using enumeration methods with improved analytical recovery, increasing sample volumes, or collecting internal seed recovery data may not substantially improve the precision of concentration estimates. The Monte Carlo simulation tool that is presented herein facilitates an analysis of experimental design to determine whether the costs and added complexity of a particular strategy to improve the precision of concentration estimates are worthwhile.

Knowledge of the impacts of measurement errors upon estimated pathogen concentrations was applied to evaluate the United States' Long Term 2 Enhanced Surface Water Treatment Rule. The prescribed treatment requirements in the regulation are based upon the average oocyst count per unit volume (without consideration of measurement errors). Analytical recovery is not addressed in this calculation because the resulting bias is assumed to counteract the bias resulting from enumeration of non-infectious oocysts. A potential flaw in this regulation is that public water systems that achieve higher analytical recovery in enumerating oocysts (due to water quality attributes or laboratory proficiency) will be more likely to observe any oocysts that are present (and to subsequently incur prescribed treatment requirements) than public water systems that achieve lower analytical recovery. Consequently, the regulation can be viewed as being less strict upon public water systems that achieve lower analytical recovery. It is therefore concluded that future rounds of monitoring in the LT2ESWTR should address analytical recovery (even if infectivity cannot yet be quantitatively addressed) in the calculation of average concentrations. If regulations were to address measurement errors in pathogen enumeration data in the future, then statistical analysis methods such as the Bayesian method presented herein could be used to evaluate the mean and variance of the pathogen concentration (and the uncertainty therein).

This chapter addressed the question of whether or not concentration data are being handled correctly in quantitative microbial risk assessment. It was demonstrated that it is inappropriate to fit distributions to sets of concentration estimates because the estimates will include some variability that is due to measurement errors (*i.e.* that is not due to variability in concentration itself). Even if a Monte Carlo QMRA is based upon a concentration distribution that has a superb fit to concentration estimates and if the QMRA addresses the uncertainty in the hyperparameters of the concentration distribution, it will not correctly describe temporal concentration variability because the concentration estimates upon which it is based are imprecise. This is even true if all data are based upon equal sample volumes, if the enumeration method consistently yielded 100% analytical recovery, and if none of the concentration estimates were manipulated (*e.g.* non-detects changed to other values) by the analyst. Fitting distributions to sets of temporally distributed concentration estimates only describes temporal variability in concentration correctly if each concentration estimate is an exact measurement of the actual concentration. Any measurement error will make concentration estimates more variable than concentration itself and unrealistic high concentrations will result in over-estimated risk. Regarding analytical recovery as an independent random variable (or even as a dependent random variable by addressing correlation) adds to this excessive variability rather than removing the fraction of the variability that is due to measurement errors. Over-estimating risk may be regarded as conservative, but it is inappropriate to use such biased QMRA approaches as a tool for regulatory comparison and evaluation of drinking water supplies. The degree of bias is case-specific and false risk assessments may indicate a need for costly treatment enhancements that are not actually necessary. Accordingly, statistical analyses such as the probabilistic models and Bayesian analyses described herein must be used to separate measurement error from temporal concentration variability (unless the measurement error has been demonstrated to be negligible) if the QMRA process is going to properly quantify risks. Likewise, measurement errors associated with infectivity, log-reduction of treatment processes, and dose-response data must also be addressed in Monte Carlo QMRA.

Chapter 7

Interpretation of Non-detect Data and Evaluation of the Sensitivity of Enumeration-based Detection Methods

In an enumeration-based detection method, there are many ways in which a non-detect result can occur: (1) the analyte particle is not present in the source, (2) the analyte particle is present in the source, but none were collected in the sample, (3) analyte particles were collected in the sample, but were lost during sample processing, or (4) analyte particles were present in the prepared sample but were not observed by the analyst. The concentration estimate associated with a non-detect sample in water is zero analyte particles per unit volume, even if the result is divided by the analytical recovery of the enumeration method to remove bias. A non-detect result in a sample that may represent a small fraction of the source, or that may have been enumerated using a method with imperfect analytical recovery, is not sufficient information to assert that the concentration is zero. Accordingly, non-detect results are often viewed with much greater scepticism than any other count of microorganisms or discrete particles.

The objectives of this chapter are as follows:

- to evaluate alternative strategies that may be used to report and interpret non-detect results (Section 7.1), and
- to estimate the sensitivity of an enumeration-based detection method as a function of known measurement errors (Section 7.2).

The chapter is concluded with a discussion of practical implications (Section 7.3).

7.1 INTERPRETATION OF NON-DETECT DATA

Non-detect enumeration data are common in many applications. Indeed, non-detect data are a favourable result in many applications such as the quantification of pathogens in treated drinking water. It is generally understood, however, that a non-detect result does not confirm that the analyte particles are absent in the source, and it is assumed that the sample collection and analysis have simply failed to detect the analyte particles that are present. This notion of uncertainty, together with a variety of mathematical issues associated with zeros, has caused non-detect results to receive considerable attention in many applications. This section summarizes issues associated with non-detect data and the method detection limit of enumeration methods (Section 7.1.1) and implements the Bayesian methods used in Chapter 5 to evaluate uncertainty in concentration and log-reduction estimates when non-detect results are obtained (Section 7.1.2).

7.1.1 Issues Associated with Non-detect Data

In many applications associated with the enumeration of microorganisms in water, non-detect data have received considerable attention. Apart from awareness that counts of zero lead to uncertain concentration estimates, three general problems arise: (1) values of zero are incompatible with logarithmic transformations, (2) many distributions do not provide a very good fit to non-detect results, and (3) concentration estimates associated with non-detect samples are often reported as being less than the method detection limit (which is not a numeric value and therefore leads to complicated data analysis).

Logarithmic transformations are used in a variety of applications associated with enumeration data. Logarithmic transformation often enables better graphical displays of results (*e.g.* Figures 6.1, 6.2, 6.7, 6.8, 6.9, and 6.10), but is not compatible with values of zero. Accordingly, some positive value must be assigned to zeros so that they can be plotted. In many applications, the geometric mean (*GM*) is calculated using logarithms of the data $\{x_i\}$ as shown in Equation 7.1. The geometric mean is used as a summary statistic in many applications (*e.g.* Christian and Pipes, 1983; Parks and VanBriesen, 2009), as an estimate of the mean of a log-normal distribution (*e.g.* Pipes *et al.*, 1977; Christian and Pipes, 1983), and is commonly used (or at least considered) as a point estimate of microorganism concentrations in quantitative microbial risk assessment (*e.g.* Regli, *et al.*, 1991; Rose and Gerba, 1991; Rose *et al.*, 1991; Haas, 1996; Benke and Hamilton, 2008). Use of the geometric mean is problematic, however, if any of the data are zero because the geometric mean cannot be calculated using Equation 7.1 and is zero if the conventional formula (that does not use logarithms) is used. In order to use the geometric mean, zeros have to be manipulated into positive values.

$$GM = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln(x_i)\right) \tag{7.1}$$

> "A problem arises if any one of the measurements is below the detection limit and is considered to be zero; the geometric mean then becomes zero. To correct this problem, any measurement below the detection limit is given the value of 1 (log 1 = 0), in effect making the calculation a modified geometric mean." – Regli *et al.*, (1991).

The second problem is how to address non-detects when fitting distributions to data. In general, continuous distributions that are used for concentrations are incompatible with values of zero (because the domain of the random variable is all positive real numbers). Accordingly, researchers who have fitted distributions to concentration estimates have had to give special treatment to non-detects (*e.g.* Jaidi, *et al.*, 2009). Non-detects have also been given special treatment when fitting distributions to counts. The Poisson-plus-added-zeros distribution has been considered (*e.g.* Christian and Pipes, 1983; Medema *et al.*, 1995; Atherholt and Korn, 1999) because non-detects may occur with a greater frequency than what is supported by the best-fit Poisson distribution.

The detection limit of an enumeration method is the lowest non-zero value that could be obtained, and is accordingly equal to one analyte particle in the volume of water that was analyzed. The Information Collection Rule Lab Manual (USEPA, 1996), for example, required non-detect results to be reported in terms of the detection limit (in cysts or oocysts per 100 L) using the formula $(<1) \cdot 100 / FVR$, in which V is the initial sample volume (in litres) and the parameters F and R are fractions of the sample that were submitted to various steps in the enumeration method. Crockett and Haas (1995) discussed various problems associated with non-detect data including analysis of datasets in which some of the values are reported as less than the detection limit. Parkhurst and Stern (1998) addressed the averaging of datasets containing zeros and concluded that zeros must be maintained as zeros when calculating the arithmetic mean to avoid bias. In evaluating the effect of averaging datasets in which non-detect (ND) samples were assigned a count of 1 (*i.e.* the detection limit), they raised the following argument.

> "When the oocyst concentration in the bulk water is very low, sampling will produce many ND samples, and the degree of overestimation will be great. On the other hand, when high concentrations predominate, ND samples become rare, and the tendency to overestimate disappears. Thus, the method overestimates at low concentrations that may be of negligible concern but fails progressively to do so at those higher concentrations that might pose a health risk." – Parkhurst and Stern (1998)

Despite this argument, manipulation of non-detect results into positive concentration values is still a common practice and detection limits of enumeration methods (which have little practical meaning as will be demonstrated in Section 7.1.2) continue to be discussed and reported. A detailed review of the literature to investigate the evolution of the method detection limit and its use, and the various ways in which non-detect data have been analyzed, is beyond the scope of this thesis, but further research in this area is clearly warranted.

7.1.2 Estimating Concentration and log-Reduction from Non-detect Data

There is general agreement that concentration estimates associated with non-detect data are uncertain, but approaches to evaluate this uncertainty have not been standardized. Often, the uncertainty is indicated using the method detection limit by reporting that the concentration is less than 1 analyte particle per volume analyzed. In Section 5.3, a Bayesian approach to quantify uncertainty in concentration estimates based upon the use of probabilistic models was presented. Similarly, a Bayesian approach was presented in Section 5.4.3 to quantify uncertainty in a log-reduction estimate calculated from two paired concentration estimates. This section addresses the use of these quantitative analysis approaches when non-detect data are obtained and discusses the results with respect to the method detection limit.

It has already been demonstrated in Section 5.3.3 (*e.g.* Figure 5.8) that the presented Bayesian approach used to estimate uncertainty in microorganism and discrete particle concentrations could be used to analyze non-detect data. The analysis used information about the analytical recovery of the enumeration method and a hypothetical datum of zero microorganisms in 1 L. The method detection limit for this scenario would be 1.0 microorganisms/L, which would often be reported as < 1 microorganism/L (or extrapolated to < 100 microoganisms/100 L). A reported concentration estimate of < 1 microorganism/L implies that the concentration in the source is actually less than 1 microorganism/L. The Bayesian analysis produces a posterior distribution that describes what the concentration might actually be given the data, knowledge about the measurement errors of the enumeration method, and a relatively uninformative prior. In both models that were used (the beta-Poisson and the negative binomial) the posterior probability that the concentration is greater than the method detection limit of 1 microorganism/L is approximately 93.5%. Accordingly, it is disingenuous to assert that the actual concentration in the source is < 1 microorganism/L.

This argument can be expanded by considering the best-case scenario in which samples are counted with 100% analytical recovery. The resulting probabilistic model for the count is $x \sim POISSON(cV)$. If a count of zero is obtained, then the likelihood is $L(c) = e^{-cV}$. If a semi-infinite uniform prior is used for concentration, then the posterior is the exponential distribution $f(c) = V \cdot e^{-cV}$. The associated cumulative posterior distribution function is $F(c) = 1 - e^{-cV}$, and the posterior probability that the concentration is greater than the method detection limit ($1/V$) is $1 - F(1/V) = e^{-1} \approx 0.368$. Therefore, the posterior probability that the concentration is greater than the detection limit is 36.8% (regardless of sample volume) if a Poisson model is assumed for random sampling error. This probability would be greater (as illustrated in the preceding example) if analytical recovery is typically below 100%. It is therefore concluded that it is inappropriate to assert that the concentration is less than the method detection limit on the basis of a single non-detect sample.

Estimation of the log-reduction of a treatment process is complicated when analyte particles are not observed in the post-treatment water and is futile if analyte particles are not observed in the pre-treatment water. A hypothetical scenario is considered herein to quantify uncertainty in log-reduction when no analyte particles are observed in the post-treatment water. The hypothetical data and parameter values that are used are summarized in Table 7.1. The table also shows the concentration and log-reduction estimates that would result if it were assumed that the non-detect post-treatment sample yielded a count of < 1. The detection limit of the post-treatment sample (having divided the assumed count of 1 by mean analytical recovery) is < 27.09 microorganisms/L and the associated estimate of log-reduction is > 3.891. The data were analyzed using Bayes' theorem as per the method of Section 5.4.3 (results not shown). The calculated posterior probability that the post-treatment concentration is above 27.09 microorganisms/L is 38.8%. The calculated posterior probability that the log-reduction is below 3.891 is 38.3%. Credible intervals might provide a more informative indication of the concentration and log-reduction values. The 95% credible interval for the post-treatment concentration is < 89.06 microorganisms/L, and the 95% credible interval for the log-reduction is > 3.361. Once again, Bayesian analysis enables a more informative assessment of the data than point-estimation approaches that replace zeros with counts of < 1.

Table 7.1: Log-reduction data and estimation with a non-detect post-treatment count

	Pre-treatment Sample	Post-treatment Sample
Count (microorganisms)	293	0 (< 1)
Volume (L)	0.005	0.1
a	18.78	24.75
b	48.75	42.29
Concentration Estimate* **(microorganisms/L)**	2.107×10^5	0 (< 27.09)
Log_{10}-reduction Estimate	(> 3.891)	

* Concentration estimates are adjusted for mean analytical recovery.

The probabilistic models and data analysis approaches that are presented in this thesis have a possible flaw with respect to the analysis of non-detect data. The models assume that the concentration can never be zero (because the mean of a Poisson distribution is necessarily positive). Crainiceanu *et al.* (2003) addressed this concern in their similar probabilistic and Bayesian models: "A basic assumption [of the model] is that the pathogen concentration in natural waters can be very small but never exactly zero". This assumption seems valid because it is likely infeasible in most situations to prove that the concentration of microorganisms or discrete particles in water is actually zero. To do so, it would be necessary to enumerate the entire source using an enumeration method that has no losses. Assuming that the concentration is not actually zero is conservative in many applications, so the models seem valid.

7.2 CASE STUDY: EVALUATING THE SENSITIVITY OF AN ENUMERATION-BASED DETECTION METHOD

In many medical and epidemiological applications, the count of a specific type of microorganism in a sample is not important as much as their presence. Accordingly, the process of collecting and analyzing a sample in these applications can be referred to as a presence-absence test. Errors in such a test can be described as diagnostic errors rather than measurement errors, because the result is a Boolean diagnosis (*i.e.* presence or absence) rather than a numeric value. The results of such a test can be categorized as true-positives, false-positives (*i.e.* a positive result when the target microorganism is absent), true-negatives, and false-negatives (*i.e.* failure to detect target microorganisms that are present). These errors are intrinsically linked to the analytical recovery of the method: analytical errors that cause microorganisms to go unobserved can lead to false-negatives, while false-positives can arise from sample contamination or false-positive counting errors. Sensitivity is defined as the probability of detection when the microorganism is actually present, and specificity is defined as the probability of non-detection when the microorganism in not actually present. These concepts have been raised in relation to enumeration methods used to detect microorganisms in water (*e.g.* Allen *et al.*, 2000; Haas, 2002).

The probabilistic models described within this thesis can be used to evaluate the sensitivity of enumeration-based detection methods in various situations if information is available about analytical recovery. This section provides a case-study analysis of the sensitivity of a method that is used to detect *Schistosoma mansoni* eggs in human feces (Teixeira *et al.*, 2007). This method has been discussed previously in Section 4.4.4.

The conventional approach that is used to evaluate the sensitivity of a detection method consists of seeding a number of samples with a specific number of the analyte particles and then evaluating the fraction of the samples that yielded non-detect results. This empirical approach is very inefficient because large numbers of samples would need to be evaluated to obtain a precise estimate of sensitivity, and the results would not apply to samples with any other quantity of seeded analyte particles. This approach was used to evaluate sensitivity for 7 different quantities of *Schistosoma mansoni* eggs in Teixeira *et al.*, (2007), as illustrated in Table 4.6. It was concluded that the numbers of seeded eggs reported in Table 4.6 were relatively precisely known (Graeff-Teixeira, personal communication). Accordingly, it was determined that the count data could be analyzed using the beta-binomial recovery model. This analysis determined that maximum likelihood estimates of the parameters of the beta distribution (which describes variability in analytical recovery for the pooled data) are a = 3.34, b = 32.90 (Section 4.4.4). Using these parameters, the beta-binomial model (Equation 4.4) can be used to estimate sensitivity for any number of eggs (n) per sample as shown in Equation 7.2 (assuming that the parameters are representative of the analytical recovery of any number of eggs considered).

$$P(X > 0 \mid n, a, b) = 1 - \frac{\Gamma(a+b)}{\Gamma(b)} \cdot \frac{\Gamma(n+b)}{\Gamma(n+a+b)} \tag{7.2}$$

While maximum likelihood estimation of the recovery distribution parameters using an appropriate probabilistic model yields good point estimates, point estimates do not address uncertainty. Uncertainty in the parameters a,b will result in uncertainty in the calculated sensitivity values. In Section 4.4.4, uncertainty in these parameters was addressed using a Bayesian Gibbs sampling approach with the prior $g(a,b) = (ab)^{-1}$. This approach is expanded herein to calculate the posterior distribution of sensitivity for various numbers of eggs. In each of 30,000 iterations, the sensitivity for each considered number of eggs (from 1 to 60) was computed using Equation 7.2 and the updated values of a,b. Following the analysis, the sequence of computed sensitivity values for each number of eggs (which was assumed to be representative of the associated posterior distribution) was evaluated to compute the posterior mean and 95% credible intervals for sensitivity.

As described above, an empirical sensitivity estimate (*i.e.* the fraction of a number of seeded samples that yield non-detect results) is an imprecise estimate of sensitivity. To illustrate this, prediction intervals for empirical sensitivity estimates (assuming r = 10 samples per number of eggs) were computed. In each Gibbs sampling iteration (and for each number of eggs), ten count values were simulated using Equation 4.4 and the updated values of a,b to compute a simulated empirical sensitivity estimate. Following the analysis, the sequence of simulated empirical sensitivity estimates for each number of eggs (which was assumed to be representative of the associated posterior distribution) was evaluated to compute 95% credible intervals. To distinguish between the two types of credible intervals, the latter is called the 95% prediction interval of the empirical sensitivity estimates. The results of all these analyses, along with the empirical sensitivity estimates that were calculated in Table 4.6, are shown in Figure 7.1. As expected the actual empirical data fall within the 95% prediction intervals, and the 95% prediction intervals are quite wide because empirical sensitivity estimates are imprecise.

This type of analysis can be conducted to rigorously investigate sensitivity for any type of enumeration method in which counts are obtained (*i.e.* in which analytical recovery can be estimated). The results are vastly more informative than a few empirical sensitivity estimates because the analysis shows the sensitivity (and uncertainty therein) for a wide range of numbers of eggs while the empirical results only provide imprecise estimates for select quantities of eggs.

7.3 PRACTICAL IMPLICATIONS

Non-detect results are common in many applications, but a non-detect sample is not sufficient evidence that the analyte particles are absent in the source. Random errors associated with sample collection, sample processing, and enumeration can cause non-detects when the analyte particles are actually present in the source. Non-detect samples have received considerable attention in many applications and they have been interpreted in many different ways. The quantitative analysis approaches presented in this thesis provide a framework to rigorously quantify the uncertainty in microorganism and discrete particle concentrations in water when non-detect results are obtained. These types of analyses should be adopted to quantify the uncertainty in concentration estimates (or

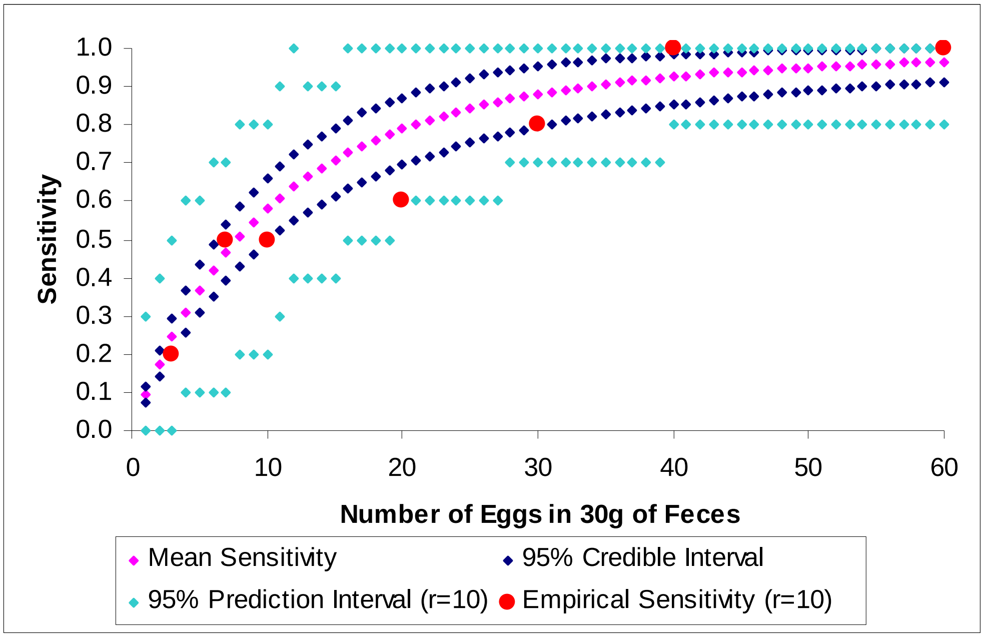


Figure 7.1: Bayesian analysis of the sensitivity of the helmintex method

log-reduction estimates when non-detect results are obtained post-treatment). These analyses depend upon reliable information about the analytical recovery of the enumeration method and appropriate quantitative analysis of the available recovery data, which have also been advocated in this thesis. The results of these analyses were compared with the common approach of providing the method detection limit of the enumeration method. It was found that reporting the method detection limit is misleading because the posterior probability that the concentration is greater than the method detection limit is substantial. The counts and concentration estimates associated with non-detect samples should be reported as zeros and not as misleading method detection limits.

The sensitivity of an enumeration method is the probability that at least one of the analyte particles that are present will be observed and this is linked to the analytical recovery of the method. Sensitivity of such methods is evaluated empirically by seeding samples with specific quantities of analyte particles and determining the fraction of such samples that yield positive counts. These results yield imprecise estimates of sensitivity and ignore the information that count data provide about analytical recovery. A Bayesian data analysis approach was presented herein that addressed the variability in analytical recovery (and the uncertainty therein) and used this information to quantify the sensitivity of the enumeration method (and uncertainty therein). This type of approach should be adopted when sensitivity of enumeration methods is evaluated because it makes use of all of the available data and provides much more information than the uncertain sensitivity estimates (*i.e.* relative frequencies of non-detects) that are obtained from the data.

Chapter 8 Variance Decomposition: A Strategy to Evaluate Sources of Error in Enumeration Data

The concept of random errors is used throughout this thesis to describe variability in count data as a function of known parameters and the parameter(s) that the analyst wishes to investigate. Knowledge of which errors contribute the most to overall variability in enumeration data is helpful because such information can aid experimental design (*e.g.* how to collect more precise recovery and concentration estimates). Throughout this thesis, Monte Carlo simulation has been used to demonstrate the impacts of various sources of variability upon enumeration data. The results have been used to facilitate experimental design recommendations and to demonstrate various kinds of bias (*e.g.* the bias of concentration estimates that are not adjusted by analytical recovery or the bias in estimates of the variability in analytical recovery or temporal concentration variability due to measurement errors). An alternative approach to these types of investigations is to use mathematical expectation and variance decomposition (Section 3.3.2).

The objectives of this chapter are as follows:

- to evaluate the factors that affect variability in analytical recovery estimates using variance decomposition and to provide recommendations for the design of recovery experiments on the basis of these results (Section 8.1), and
- to evaluate the factors that affect variability (and bias) in concentration estimates using variance decomposition and to provide experimental design recommendations on the basis of these results (Section 8.2).

The chapter is concluded with a discussion of practical implications (Section 8.3).

8.1 VARIANCE DECOMPOSITION OF PROBABILISTIC RECOVERY MODELS

This section addresses variance decomposition of probabilistic recovery models such as the beta-binomial, beta-Poisson, and negative binomial (and their partial sample analysis counter-parts) that were presented in Section 4.2. The variance decomposition results are used (1) to show that the variance of analytical recovery estimates is a biased estimate of the variance of analytical recovery, (2) to aid experimental design of recovery experiments (*e.g.* the choice of seed dose, seeding method, and number of replicate samples), and (3) to develop a method of moments parameter estimation approach that addresses the measurement errors in the recovery estimates.

8.1.1 Analytical Recovery Modelled as a Probability

Analytical recovery has often been modelled as a probability that each particle in a sample will be observed. If each of the *n* particles in a sample has equal probability (*p*) of being observed, then the resulting number of observations (*x*) will follow the binomial distribution $x \sim BINOMIAL(n,p)$. The mean and variance of *x* are $E[x] = np$ and $Var[x] = np \cdot (1 - p)$ respectively, and it follows that $E[x/n] = p$. Therefore, in the context of estimating analytical recovery, the number of seeded particles that are observed (*x*) divided by the number of seeded particles (*n*) is an unbiased estimate of the probability of recovery (*p*).

Known number of seeded particles with variable analytical recovery

This section applies variance decomposition to this binomial recovery model when the probability of recovery varies randomly among samples (*i.e.* when there is non-constant analytical recovery) with mean μ_p and variance σ_p^2. This is analogous to the beta-binomial recovery model presented in Section 4.2.2 (in which the number of seeded particles is precisely known) except that it is generalized for any non-constant analytical recovery distribution on the interval $0 < p < 1$. The mean of the recovery estimates is $E[x/n] = \mu_p$, which shows that the number of observed particles divided by the number of seeded particles is an unbiased estimate of mean analytical recovery. The variance of these recovery estimates is shown in Equation 8.1. Derivation of these equations is provided in Appendix E. Equation 8.1 shows that the variance of the recovery estimates is a biased estimate of the variability in analytical recovery itself whenever the sum of the first two terms is not equal to zero, and that the bias diminishes as the number of seeded particles is increased. If non-constant analytical recovery is beta-distributed, then the first term is always positive (see proof in Appendix E) and the variability of these recovery estimates is always an over-estimate of the variability in analytical recovery itself.

$$Var\left[\frac{x}{n}\right] = \frac{\mu_p}{n} - \frac{\mu_p^2 + \sigma_p^2}{n} + \sigma_p^2 \tag{8.1}$$

Variable number of seeded particles with variable analytical recovery

This section further applies variance decomposition to account for the situation in which the number of seeded particles also varies randomly among samples (*i.e.* there is seeding error) with mean μ_n and variance σ_n^2. In this case, $E[x/\mu_n] = \mu_p$, so the unbiased recovery estimate is the count divided by the mean number of seeded particles. The associated variance is shown in Equation 8.2.

$$Var\left[\frac{x}{\mu_n}\right] = \frac{\mu_p}{\mu_n} - \frac{\left(\mu_p^2 + \sigma_p^2\right)}{\mu_n} + \frac{\left(\mu_p^2 + \sigma_p^2\right)\sigma_n^2}{\mu_n^2} + \sigma_p^2 \tag{8.2}$$

Poisson-distributed seed doses can arise when samples are seeded with a volume (*V*) of a stock with known concentration (*c*): the resulting dose is $\lambda = cV$ (Section 4.2.3) and $\mu_n = \lambda$ and $\sigma_n^2 = \lambda$. Substitution of these values into Equation 8.2 yields Equation 8.3. This equation shows that the variance of the recovery estimates is an over-estimate of the variability in analytical recovery itself and that the bias diminishes as the expected number of seeded particles (λ) is increased.

$$Var\left[\frac{x}{\lambda}\right] = \frac{\mu_p}{\lambda} + \sigma_p^2 \qquad (8.3)$$

These equations are used to investigate the experimental design of recovery experiments in Section 8.1.4, and to investigate estimation of non-constant analytical recovery distribution parameters using the method of moments in Section 8.1.5. The variance decomposition could be expanded further to address the impacts of uncertainty in μ_n, σ_n^2, or λ, but this is not addressed herein.

8.1.2 Analytical Recovery Modelled as a Rate

In the negative binomial recovery model (Section 4.2.4), analytical recovery was modelled as a rate (p) that reduced the number of observed analyte particles (x) in a sample with a Poisson-distributed quantity of seeded particles (λ). The resulting model is $x \sim POISSON(\lambda p)$, which has mean $E[x] = \lambda p$ and variance $Var[x] = \lambda p$. In the context of estimating analytical recovery, the number of seeded particles that are observed (x) divided by the expected number of seeded particles (λ) is an unbiased estimate of the recovery rate (p). Variance decomposition can be applied to this Poisson model to investigate the impacts of non-constant analytical recovery (with mean μ_p and variance σ_p^2) upon the variability in the recovery estimates. This approach is generalized for any non-constant analytical recovery distribution on the interval $p > 0$ (*i.e.* it is not specific to the gamma distribution that is used in the negative binomial model). The variance of the recovery estimates is $Var[x/\lambda] = \mu_p/\lambda + \sigma_p^2$. This result is the same as Equation 8.3. The reason is that the Poisson-distributed seeding error and binomially distributed analytical error in the formulation of Equation 8.3 is equivalent to the Poisson distribution $x \sim POISSON(\lambda p)$ as shown in Equation 4.6. Accordingly, the results of variance decomposition are the same (except that analytical recovery is a probability in one and a rate that may exceed 100% in the other).

8.1.3 Partial Sample Analysis in Recovery Experiments

Partial sample analysis addresses the situation in which only a fraction (θ) of a seeded sample is enumerated. As discussed in Section 4.2.5, partial sample analysis can be incorporated into the binomial model for analytical error as $x \sim BINOMIAL(n,\theta p)$, and can be incorporated into the Poisson model for seeding and analytical error as $x \sim POISSON(\lambda\theta p)$. Both models assume that the sample was well-mixed prior to any sub-sampling. In the former case (which is related to the beta-binomial recovery model), $E[x/n\theta] = p$, so $x/n\theta$ is the unbiased recovery estimate. In the latter case (which is related to the beta-Poisson and negative binomial recovery models), $E[x/\lambda\theta] = p$, so $x/\lambda\theta$ is the unbiased recovery estimate. This section addresses evaluation of the variability of these unbiased recovery estimates when analytical recovery varies with mean μ_p and variance σ_p^2. Using variance decomposition, the variance of the unbiased recovery estimates for the binomial model is shown in Equation 8.4, and the variance of the unbiased recovery estimates for the Poisson model is shown in Equation 8.5. Derivation of these equations is provided in Appendix E. The variance decomposition could be expanded further to address the impacts of uncertainty in θ, but this is not addressed herein.

$$Var\left[\frac{x}{n\theta}\right] = \frac{\mu_p}{n\theta} - \frac{\mu_p^2 + \sigma_p^2}{n} + \sigma_p^2 \tag{8.4}$$

$$Var\left[\frac{x}{\lambda\theta}\right] = \frac{\mu_p}{\lambda\theta} + \sigma_p^2 \tag{8.5}$$

In Equation 8.4, $n\theta$ is the ‘equivalent seed dose’. If the equivalent dose is substituted into Equation 8.1, the result is different than Equation 8.4. Therefore, when the number of seeded particles is precisely known, partial sample analysis has an effect upon measurement error beyond making the effective dose smaller than the actual dose. As described in Section 4.3, the number of seeded particles in the sample is precisely known, but the number of seeded particles in the sub-sample is random (*i.e.* there is a sub-sampling measurement error associated with partial sample analysis). The effect of this sub-sampling error can be evaluated by subtracting Equation 8.1 (with the equivalent seed dose $n\theta$) from Equation 8.4, and the result in Equation 8.6. This additional measurement error diminishes as θ approaches 100%, as $\mu_p^2 + \sigma_p^2$ approaches zero, and as n increases. Substitution of the equivalent seed dose $\lambda\theta$ into Equation 8.3 yields the same result as Equation 8.5, so partial sample analysis only affects measurement error by changing the effective seed dose in this type of model.

$$\Delta Var\left[\frac{x}{n\theta}\right] = \left(\frac{1}{\theta} - 1\right)\frac{\left(\mu_p^2 + \sigma_p^2\right)}{n} \tag{8.6}$$

8.1.4 Using Variance Decomposition to Aid Recovery Experiment Design

The objective of recovery experiments in which many samples are seeded and analyzed under similar conditions is to evaluate the variability in analytical recovery among samples (*e.g.* to estimate the mean and standard deviation of analytical recovery). The conventional recovery estimate is the number of observed particles divided by the presumed number of seeded particles. Monte Carlo simulation was used in Section 4.3 to show how the mean and variance of these recovery estimates was related to the actual mean and variance of analytical recovery. Numerous experimental design scenarios were compared, and the results were used to infer what sources of error were contributing to the error in recovery estimates, and what changes in experimental design could enable better estimation of the mean and standard deviation of analytical recovery from the recovery estimates.

In this chapter, variance decomposition has been used to evaluate the variance of recovery estimates as a function of the parameters in the probabilistic recovery models (which circumvents the need for Monte Carlo simulation). Monte Carlo simulation is useful to evaluate the distribution of possible counts (or functions thereof), but is not necessary to evaluate variability. It is evident in Equations 8.1 through 8.5 that the variance of the unbiased recovery estimates is greater than the variability in analytical recovery itself (except for certain circumstances in Equations 8.1 and 8.4). The objective of recovery experiment design should be to seed samples such that each sample’s analytical recovery can be estimated precisely, and also to process enough samples to enable estimation of the mean and variance of recovery with good precision. The following analyses based upon variance decomposition provide a simple, yet rigorous, mathematical assessment of these issues.

Choosing an appropriate seed dose

In Equations 8.1 through 8.5, the variance of the recovery estimates is made up of two components: the variability in analytical recovery and a term (or a collection of terms) that summarizes the measurement errors associated with estimating analytical recovery. The measurement error causes the variability of the recovery estimates to be greater than the variability in analytical recovery itself. Even if estimating the variability in analytical recovery is not the objective of a recovery experiment, this bias is important because it will affect the uncertainty in the mean analytical recovery. The quantity of seeded particles per sample is an experimental design parameter that affects the magnitude of the measurement errors, so the seed dose should be selected such that measurement errors are small. The analysis herein assumes that analytical recovery is independent of seed dose. If, however, the seed dose affects the analytical recovery of an enumeration method in a way that would make the results unrepresentative of the application for which the data are being collected, then representative seed doses should be used regardless of measurement error.

If the measurement errors (σ_e^2) associated with a particular seed dose are trivially small compared to the variability in analytical recovery (σ_p^2), then the total variability in the recovery estimates (σ_t^2) will accurately describe the variability in analytical recovery. Conversely, large measurement errors cause the variability of the recovery estimates to be more indicative of measurement error than variability in analytical recovery itself. Variance decomposition can be used to compute the magnitude of the measurement errors relative to the variability in analytical recovery given specific values of μ_p, σ_p^2, and the seed dose. To use these as an experimental design tool, prior estimates of μ_p and σ_p^2 must be available (*e.g.* from preliminary results or similar experiments). For any given seed dose, the variance decomposition equations can then be used to evaluate how much (*e.g.* k%) of the total variability in hypothetical recovery estimates would be due to measurement error.

Herein, a maximum value of k is specified to determine the minimum seed dose that would be required. The result associated with Equation 8.1 (which is related to the beta-binomial recovery model in which the number of seeded particles, n, is precisely known) is Equation 8.7. Similarly, The result associated with Equation 8.3 (which is related to the beta-Poisson and negative binomial recovery models in which the seed doses are Poisson-distributed with mean λ) is Equation 8.8.

$$n > \frac{100-k}{k}\left[\frac{\mu_p\left(1-\mu_p\right)}{\sigma_p^2}-1\right] \tag{8.7}$$

$$\lambda > \frac{100-k}{k}\left(\frac{\mu_p}{\sigma_p^2}\right) \tag{8.8}$$

To illustrate use of this experimental design tool, two scenarios are compared herein: a situation in which analytical recovery is low and highly variable (Scenario 1), and a situation in which analytical recovery is high with low variability (Scenario 2). The mean and standard deviation of analytical recovery in each scenario, as well as the minimum seed doses calculated using Equations 8.7 and 8.8 (with k = 50%) are summarized in Table 8.1. These scenarios are based upon example beta distributions that were plotted in Figure 4.4 and subsequently used in Monte Carlo experimental design analyses. In this particular application these mean and variance values do not correspond to a specific type of distribution. The value k = 50% was selected in this analysis because it is the point at which the measurement error of the recovery estimates and the variability in analytical recovery contribute equally to overall variability in the recovery estimates. For any smaller seed dose, variability in the recovery estimates would be more indicative of measurement error than variability in analytical recovery.

Table 8.1: Calculation of minimum seed doses in recovery experiments

	Scenario 1	**Scenario 2**
μ_p	0.2529	0.7518
σ_p	0.0854	0.0435
Minimum value of *n* (*k* = 50%)	25	98
Minimum value of λ (*k* = 50%)	34.68	396.69

Comparison of the minimum seed doses shows that higher seed doses are needed in Scenario 2 than in Scenario 1 to overcome the effects of measurement error in the recovery estimates. This is because measurement error has a greater proportional impact upon the total variability in recovery estimates when the variability in analytical recovery is small. In each scenario, the minimum seed dose (λ) that is needed when the number of seeded particles is Poisson-distributed is greater than the minimum precisely known number of seeded particles (n). This is because seeding error increases the measurement error in the recovery estimates. The increase in the minimum quantity of seeded particles is greatest in Scenario 2. The comparison between precisely known and Poisson-distributed seed doses is investigated in greater detail below.

Figure 8.1 shows the effect of measurement error upon total variability in recovery estimates (in terms of standard deviations) for both scenarios and for seed doses up to 500 particles. Figure 8.1a shows the results when the number of seeded particles is precisely known (using the variance decomposition result shown in Equation 8.1). Figure 8.1b shows the results when the number of seeded particles is Poisson-distributed (using the variance decomposition result shown in Equation 8.3). These figures demonstrate that the effects of measurement error upon total variability in recovery estimates diminish as the seed dose is increased: "there was a quantity of seeded particles (known or Poisson-distributed) beyond which the benefits of further increases [in the number of seeded particles] were negligible" (Schmidt *et al.*, 2010a). They also show that these effects become trivial at smaller seed doses when analytical recovery is more variable: "these transitions occurred at lower numbers of seeded particles for methods with more variable recovery" (Schmidt *et al.*, 2010a).

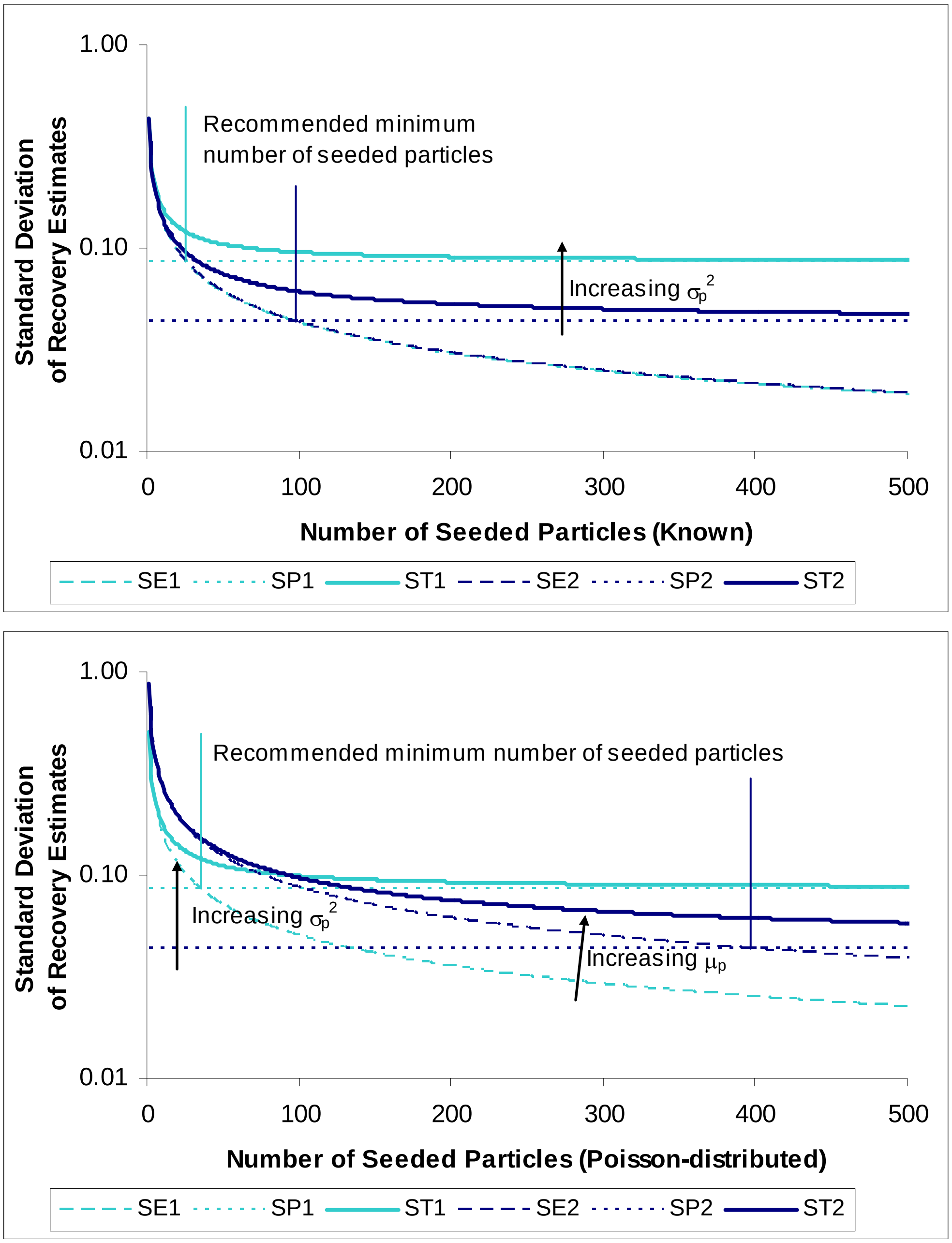


Figure 8.1: Effect of seed dose on measurement errors in analytical recovery estimates

'SE' represents the standard deviation of the measurement error (σ_e), 'SP' represents the standard deviation of analytical recovery (σ_p), and 'ST' represents the total standard deviation of the recovery estimates (σ_t). Scenario 1 uses μ_p = 0.2529, σ_p = 0.0854 and Scenario 2 uses μ_p = 0.7518, σ_p = 0.0435. The two panels compare precisely known and Poisson-distributed numbers of seeded particles.

The ‘SE1’ and ‘SE2’ series in Figure 8.1a (which represent the measurement error for the two scenarios) coincide because the term $\mu_p - \mu_p^2$ in the measurement error calculation is almost the same when $\mu_p = 0.2529$ as when $\mu_p = 0.7518$.

Comparison of the ‘SE1’ and ‘SE2’ series in Figure 8.1b shows that higher mean analytical recovery leads to higher measurement error when Poisson-distributed seed doses are used. To investigate the effects of seeding error upon measurement error in recovery estimates (without the effects of analytical error), Equation 8.1 was subtracted from Equation 8.3 (with $n = \lambda$) to yield Equation 8.9. This equation shows that the contribution of seeding error to overall measurement error in recovery estimates (ΔVar) increases as $\mu_p^2 + \sigma_p^2$ increases. Schmidt *et al.* (2010a) concluded that “the observed difference in probability intervals between precisely known and Poisson-distributed numbers of seeded particles was greater for methods with higher analytical recovery”, but this is only true if the increase in μ_p^2 is greater than a possible decrease in σ_p^2. The values of $\mu_p^2 + \sigma_p^2$ in Scenarios 1 and 2 from the preceding analysis are 0.0712 and 0.5672 respectively. This is why the increase in the calculated minimum number of seeded particles is small in Scenario 1 and is large in Scenario 2.

$$\Delta Var\left[\frac{x}{\lambda}\right] = \frac{\mu_p^2 + \sigma_p^2}{\lambda} \tag{8.9}$$

The effects of partial sample analysis in recovery experiments were addressed in Section 8.1.3. Partial sample analysis reduces the effective seed dose, and can add measurement error if the number of seeded particles is precisely known. As discussed in Section 4.3, partial sample analysis can partially negate the benefits of seeding samples with precisely known numbers of seeded particles in recovery experiments.

Choosing an appropriate number of samples

The preceding discussion addressed the precision of recovery estimates with respect to various quantities of seeded particles. In a recovery experiment that is used to quantify the mean and variance of analytical recovery, the number of samples that are analyzed will have an effect upon the sample mean and sample standard deviation of the recovery estimates. In Section 4.3 (and in Schmidt *et al.*, 2010a), Monte Carlo simulation was used to generate 95% probability intervals for the sample mean and the sample standard deviation of the recovery estimates associated with various experimental designs and hypothetical recovery parameters (a,b). These probability intervals were used as a graphical display of precision and may be helpful in selecting an appropriate number of samples in a recovery experiment.

Through the use of variance decomposition, the population mean (μ) and population variance (σ^2) of the recovery estimates has been computed as a function of μ_p and σ_p^2. According to the central limit theorem, the distribution of the sample mean ($\bar{p}$) of the recovery estimates is approximately normally distributed, $\bar{p} \sim NORMAL\left(\mu, \sigma^2/r\right)$, if the recovery estimates are approximately normally distributed or if the number of samples (r) is large. Similarly, the value $(r - 1) \cdot s_p^2$, in which s_p is the sample standard deviation of the recovery estimates, will follow a chi-squared distribution with *r*-1 degrees of freedom if the data are approximately normally distributed. If these assumptions are valid, then 95% probability intervals for the sample mean and sample standard deviation of the recovery estimates can be calculated using Equations 8.10 and 8.11.

$$\mu - \frac{z_{0.025}}{\sqrt{r}} \cdot \sigma < \overline{p} < \mu + \frac{z_{0.025}}{\sqrt{r}} \cdot \sigma \tag{8.10}$$

$$\sqrt{\frac{\chi^2_{r-1,0.975} \cdot \sigma^2}{r-1}} < s_p < \sqrt{\frac{\chi^2_{r-1,0.025} \cdot \sigma^2}{r-1}} \tag{8.11}$$

To compare the Monte Carlo simulation approach with calculation of the probability intervals, several scenarios were considered herein. Each scenario used the parameters a = 73.26 and b = 24.18 (which corresponds to μ_p = 0.7518, σ_p = 0.0435). 95% probability intervals were obtained using seed doses of 50 and 10 particles (known, or Poisson-distributed) and numbers of samples ranging from 3-20. The results are shown in Figure 8.2. These results show that the variance decomposition approach to computing the probability intervals yields quite similar results to the Monte Carlo approach. Any discrepancy between the two approaches would result from violation of the normality assumption upon which Equations 8.10 and 8.11 are based. This approach based upon variance decomposition is much easier to implement than the Monte Carlo approach and results can be calculated quickly and easily without the need for extensive simulations. Therefore, there is relatively little value in using the Monte Carlo approach that was presented in Section 4.3 and Schmidt *et al.* (2010a) unless the distribution of counts (or associated recovery estimates) departs substantially from normality. The use of variance decomposition and Equations 8.10 and 8.11 is a simple approach to evaluate the effect of the number of samples upon the estimated mean and standard deviation of analytical recovery. The number of samples in a recovery experiment should be chosen so that the computed 95% probability intervals are satisfactorily narrow (which would correspond to precise estimates of the mean and standard deviation of analytical recovery from the experimental results).

8.1.5 Using Variance Decomposition in Method of Moments Parameter Estimation

In Section 4.4.1, the method of moments was used to fit distribution parameters to sets of recovery estimates. The method was rejected as a suitable parameter estimation method because the variability of the recovery estimates is greater than the variability in analytical recovery itself due to measurement errors. Given the variance decomposition results presented herein (*e.g.* Equations 8.1 and 8.3), the portion of the total variability that is due to measurement errors can be estimated and used to correct the estimate of the variability in analytical recovery. The method of moments can then be applied to fit parameters using this corrected variance value. Given the sample mean ($\overline{p}$) and sample standard deviation (s_p^2) of the unbiased recovery estimates obtained using a precisely known number of seeded particles (n), the substitutions $\overline{p}$ = μ_p and s_p^2 = $Var[x/n]$ can be applied to Equation 8.1 to estimate σ_p^2 (which results in Equation 8.12). Similarly, the substitutions $\overline{p}$ = μ_p and

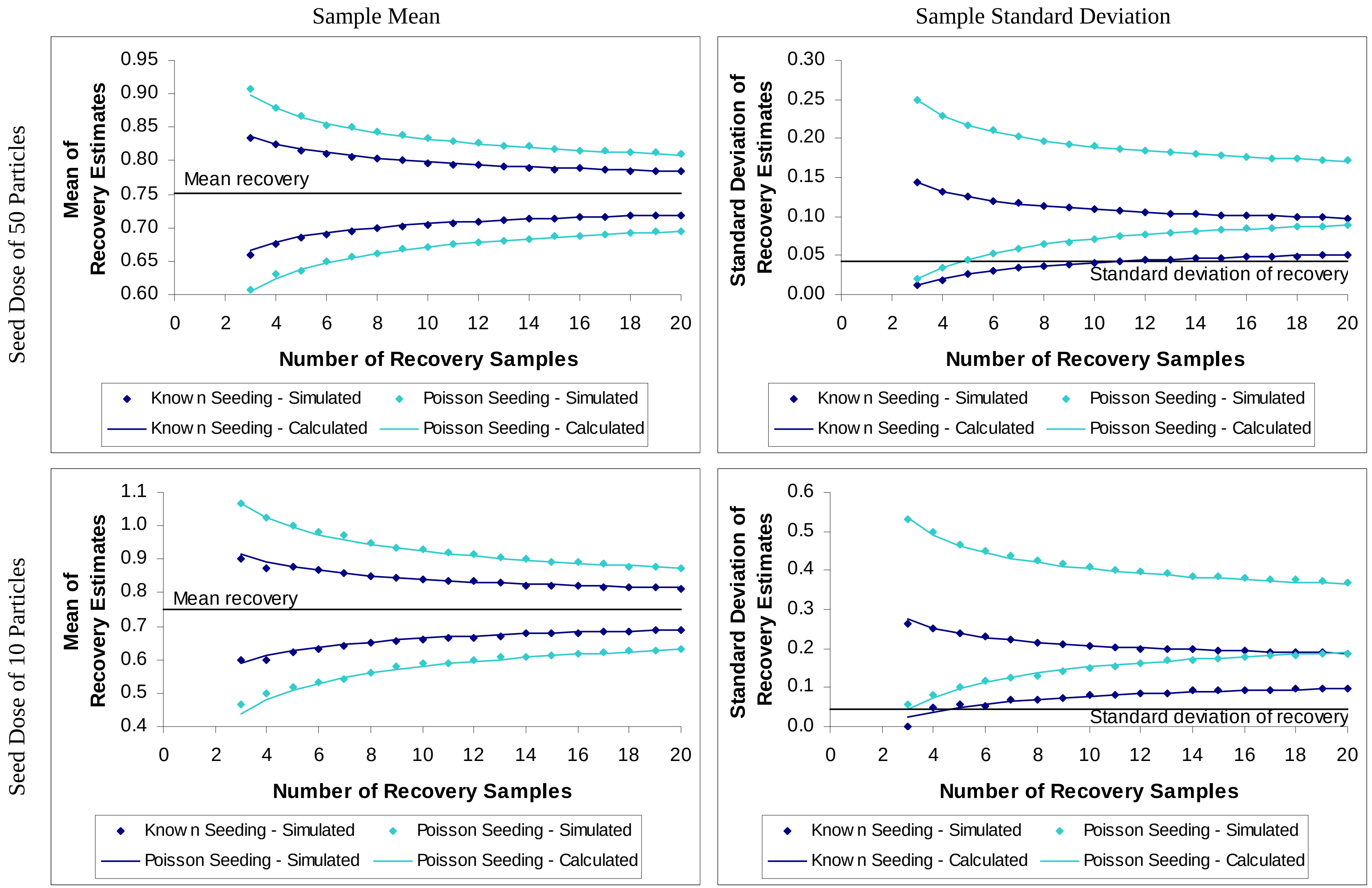


Figure 8.2: 95% probability intervals for 3-20 samples with known or Poisson seeding of 10 or 50 particles (*a*,*b* = 73.26,24.18)

$s_p^2 = Var[x/\lambda]$ can be applied to Equation 8.3 to estimate σ_p^2 when the number of seeded particles is Poisson-distributed (which results in Equation 8.13). Equations 8.12 and 8.13 can both yield negative estimates of variance depending upon the specific values of $\bar{p}$ and s_p^2 (because $\bar{p}$ is an imprecise estimate of μ_p): therefore, the method of moments will not always yield feasible parameter estimates.

$$\sigma_p^2 = \frac{n \cdot s_p^2 - \bar{p}(1-\bar{p})}{(n-1)} \tag{8.12}$$

$$\sigma_p^2 = s_p^2 - \frac{\bar{p}}{\lambda} \tag{8.13}$$

This method of moments approach was used to analyze the Table 4.4 data (using Equation 8.12 for the beta-binomial model and Equation 8.13 for the beta-Poisson and negative binomial models), and the results were compared to the analyses presented in Table 4.5. Table 8.2 shows the results of this comparison. In each case, the modified method of moments approach computed a positive value of variance and feasible parameter values that are near the maximum likelihood estimates. This may be because 9 samples seeded with 608 particles yields a relatively precise estimate of the mean. If the estimated mean is imprecise, then this method of moments approach would yield unreliable (if not infeasible) parameter estimates. Accordingly, maximum likelihood estimation is a preferable and more robust approach to estimate parameters with appropriate regard for measurement errors.

8.2 VARIANCE DECOMPOSITION OF PROBABILISTIC CONCENTRATION MODELS

This section addresses variance decomposition of probabilistic concentration models such as the beta-Poisson, and negative binomial models (and their partial sample analysis counter-parts) that were presented in Section 5.1. The variance decomposition results are used to evaluate the measurement error in individual concentration estimates as a function of sample volume and analytical recovery (Section 8.2.1), to aid experimental design by assessing alternative strategies that may enable collection of more precise concentration estimates (Section 8.2.2), and to show that the variance of

Table 8.2: Method of moments parameter estimation based on variance decomposition

	Beta-binomial	**Beta-Poisson**	**Negative Binomial**
Estimated Mean	μ_p = 0.751827	μ_p = 0.751827	μ_p = 0.751827
Estimated Variance	σ_p^2 = 0.001591	σ_p^2 = 0.000659	σ_p^2 = 0.000659
Method of Moments Parameter Estimates	a = 87.42 b = 28.86	a = 212.20 b = 70.05	α = 858.11 β = 0.000876
Maximum Likelihood Parameter Estimates	a = 104.53 b = 34.51	a = 287.08 b = 94.76	α = 1236.16 β = 0.000608

concentration estimates is a biased estimate of the actual concentration variance due to measurement errors (Section 8.2.3). In these models, the number of analyte particles that are observed in a sample (*x*) depends upon the concentration in the source (*c*), the volume of sample that is enumerated (*V*), and the analytical recovery of the enumeration method (*p*). Both models are based upon the distribution $x \sim POISSON(cVp)$ in which both the mean and variance of the counts are *cVp*. It follows that $E[x/V] = cp$ and that $E[x/Vp] = c$, so the count per unit volume is a biased concentration estimate whenever analytical recovery is not 100%, and x/Vp is an unbiased estimate of concentration.

8.2.1 Variance Decomposition with Constant Concentration

This section applies variance decomposition to the model $x \sim POISSON(cVp)$ when analytical recovery varies randomly among samples with mean μ_p and variance σ_p^2. This model is consistent with the beta-Poisson and negative binomial enumeration models (Equations 5.2 and 5.4 respectively). The mean of the concentration estimate $x/V\mu_p$ is $E[x/V\mu_p] = c$, which shows that the number of observed particles divided by the mean analytical recovery of the enumeration method is an unbiased estimate of concentration. The variance of these concentration estimates is shown in Equation 8.14 (Schmidt *et al.*, 2008). Derivation of this equation is provided in Appendix E.

$$Var\left[\frac{x}{V\mu_p}\right] = \frac{c}{V\mu_p} + \frac{c^2\sigma_p^2}{\mu_p^2} \tag{8.14}$$

If analytical recovery is constant (*i.e.* if $\sigma_p^2 = 0$), then *x* is Poisson-distributed (as described above) and $Var[x/V\mu_p] = c/V\mu_p$. Therefore, the second term in Equation 8.14 represents the over-dispersion of fully replicated counts (with respect to the Poisson distribution) due to non-constant analytical recovery. This is consistent with the assertion in Section 5.2.1 that variability in analytical recovery can cause data to be over-dispersed. Equation 8.14 shows that unbiased concentration estimates with different sample volumes do not have homogeneous measurement error. This has important implications for conventional hypothesis tests and regression analyses that may be conducted upon concentration estimates: the tests are not valid unless the data are transformed so that the measurement error is uniform. If the concentration is increased in Equation 8.14, then the variability of the concentration estimates is increased. Although the variability increases as a function of concentration, the relative standard deviation (which is often regarded as a measure of precision) will decrease. This is discussed further in Section 8.2.2.

Partial sample analysis addresses the situation in which only a fraction (θ) of a sample is enumerated. As discussed in Section 5.1.3, partial sample analysis can be incorporated into the Poisson model for random sampling and analytical error as $x \sim POISSON(cV\theta p)$ if the samples are well-mixed prior to sub-sampling. If both the initial sample volume and fraction of sample that is enumerated are precisely known, then the effective volume (θV) can be substituted into Equations 8.14. The count obtained in fraction θ of a sample with volume *V* is as variable as the count obtained from enumeration of an entire sample with volume θV. Variance decomposition could be applied to address variability in θ (and/or *V*) among samples, but this is not addressed herein.

8.2.2 Using Variance Decomposition to Aid Collection of More Precise Concentration Estimates

With the exception of presence-absence assays and qualitative analyses, the purpose of collecting water samples and searching for specific types of microorganisms or discrete particles is to quantify their abundance in the water. It is generally assumed that each sample is representative of some volume of water in which the analyte particle density can be regarded as a homogeneous concentration. Accordingly, the objective should be to measure that concentration as accurately and precisely as possible. Throughout this thesis, it has been asserted that enumeration methods with imperfect analytical recovery can cause concentration estimates to be inaccurate and that there are unavoidable measurement errors that make concentration estimates imprecise even if the enumeration method has perfect analytical recovery. The bias in concentration estimates that is due to imperfect analytical recovery can easily be resolved by dividing the count per unit volume by the mean analytical recovery of the enumeration method (or a sample-specific estimate of analytical recovery). The precision of recovery estimates can only be improved by collecting samples with less analytical error or through replication (*i.e.* collecting several samples that are believed to be representative of the same concentration and averaging the results to obtain a single, more precise, concentration estimate). These types of issues have been addressed previously (following the development of appropriate probabilistic models) by comparing credible intervals for various hypothetical scenarios (Section 5.6) and by Monte Carlo simulation (*e.g.* Section 6.2). In this chapter, variance decomposition has been used to describe the variance of concentration estimates as a function of the parameters in the probabilistic models (which circumvents the need for Monte Carlo simulation). Monte Carlo simulation is only necessary to evaluate the distribution of possible counts (or functions thereof), not to evaluate overall variability.

Effect of sample volume and analytical recovery upon concentration estimates

In most situations, there is only so much that can be done to refine the analytical recovery of an enumeration method. Often, the only strategies to obtain more precise concentration estimates are (1) to increase the sample volume, (2) to obtain sample-specific recovery estimates, or (3) to average the results of several replicates. Section 5.6.2 addressed the effect of sample volume and analytical recovery upon 95% credible intervals obtained using the Bayesian analysis methods that were presented in Section 5.3. Variance decomposition, however, provides a more efficient approach to investigate the effects of sample volume and analytical recovery upon the precision of concentration estimates. For a given concentration (c), and a given mean and variance of analytical recovery (μ_p, σ_p^2), Equation 8.14 can be used to evaluate the precision of concentration estimates as a function of volume. In Equation 8.14, the first term is attributed to random sampling and analytical error because it is the measurement error that would remain if the analytical recovery of the enumeration method were constant. The second term addresses the additional measurement error in concentration estimates that is due to non-constant analytical recovery. The volume at which k% of the total variability in concentration estimates is attributed to random sampling and analytical error can be determined using Equation 8.15 (Schmidt *et al.*, 2008). It can reasonably be asserted for k > 99% that the total measurement error is predominantly due to random sampling and analytical error and that reducing the variance of analytical recovery would not improve the precision of concentration estimates. Similarly, it can reasonably be asserted for $k < 1\%$ that the total measurement error is predominantly due to non-constant analytical recovery and that increasing the sample volume would not improve the precision of concentration estimates.

$$V = \frac{100-k}{k}\left[\frac{\mu_p}{c \cdot \sigma_p^2}\right] \tag{8.15}$$

Given a target value of k, μ_p and σ_p^2 (or preliminary estimates thereof), and a hypothetical value (or range of practically relevant values) of concentration, Equations 8.15 can be used as an experimental design tool. Additionally, Equation 8.14 can be used with several alternative values of μ_p, σ_p^2 to evaluate the effects of mean analytical recovery and the variability in analytical recovery upon the precision of concentration estimates. An illustrative example of this process is provided herein (and in Schmidt *et al.,* 2008) with two alternative scenarios concerning analytical recovery: Scenario 1 is a situation in which analytical recovery is low and highly variable ($\mu_p = 0.3$, $\sigma_p^2 = 0.007$), and Scenario 2 is a situation in which analytical recovery is high with low variability ($\mu_p = 0.8$, $\sigma_p^2 = 0.002$). The concentration is set to 100 microorganisms/L in this example. Figure 8.3 shows the contributions of the two terms in Equation 8.14 to the total variability in concentration estimates. The figure shows that when the sample volume is low, the error in concentration estimates is due to random sampling and analytical error ('SAE'). In this case, the error can be reduced slightly by increasing the mean analytical recovery of the enumeration method, but not by reducing the variability in analytical recovery. The variability is reduced most effectively by increasing sample volumes. In both presented scenarios, increasing the sample volume from 10 mL to 100 mL reduces the variability of the concentration estimate by approximately an order of magnitude. As the sample volume is increased, however, the effect of non-constant analytical recovery ('NAR') increases until it eventually becomes the dominant error (after which further increases in sample volume will not reduce the variability of the concentration estimate). The transition from dominance of random sampling and analytical error to dominance of non-constant analytical recovery occurs at a lower volume when the variability in analytical recovery is higher (and also when the mean analytical recovery is lower, although this is not shown here). The volume at which the two types of error contribute equally to variability (*i.e.* when $k = 50\%$) is 0.43 L in the first scenario and 4 L in the second scenario. In both scenarios shown in Figure 8.3, the precision of the concentration estimate is not reduced by using sample volumes above 20L. These volumes would be higher if the concentration were lower.

Understanding the effect of analytical recovery upon the precision of concentration estimates is important. One reason is that collecting sample-specific internal seed recovery estimates will not improve (and may conceivably worsen) the precision of concentration estimates if non-constant analytical recovery does not contribute significantly to overall measurement error (*e.g.* Section 6.2; Schmidt and Emelko, 2010a). Additionally, this knowledge helps to determine whether or not improving the analytical recovery of the enumeration method is consequential. Considerable effort

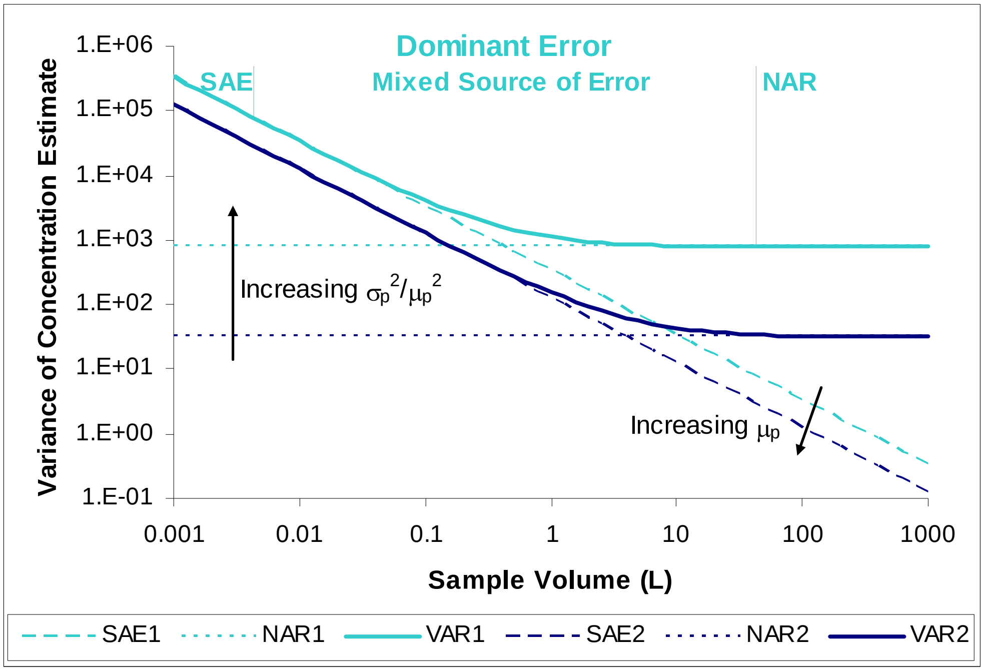


Figure 8.3: Separation of variability in concentration estimates (Schmidt *et al.*, 2008) 'SAE' represents the portion of the variance that is due to random sampling and analytical error, 'NAR' represents the portion of the variance that is due to non-constant analytical recovery, and 'VAR' represents the total variability in concentration estimates. Scenario 1 uses μ_p = 0.3, σ_p^2 = 0.007, Scenario 2 uses μ_p = 0.8, σ_p^2 = 0.002, and the concentration is c = 100 microorganisms/L.

has been put into method development research associated with the enumeration of *Cryptosporidium* and it is often implied that better concentration estimates could be obtained if the analytical recovery were improved. Variance decomposition shows that there are many situations in which increasing the mean analytical recovery or decreasing the variance of analytical recovery would not appreciably improve the precision of concentration estimates. Increasing mean analytical recovery towards 100% is generally only helpful because (1) analysts who ignore analytical recovery and report the counts per unit volume will report less biased concentration estimates, and (2) it will increase the probability of detecting at least one particle (*i.e.* the sensitivity of enumeration) when the concentration is low. If improving analytical recovery (by increasing mean analytical recovery or decreasing variability in analytical recovery) does not substantially improve the precision of concentration estimates in a particular situation, then the reverse argument would be that a method in which the analytical recovery is slightly lower or slightly more variable may not substantially compromise the precision of concentration estimates. Accordingly, method development should be aimed towards methods that enable larger volumes to be enumerated or that are logistically preferable (*e.g.* less costly, less time-consuming, or more broadly applicable)

Reducing the relative standard deviation of concentration estimates

In Equation 8.15, it is clear that as the concentration increases, so too will the variability in concentration estimates. This equation provides a description of absolute variability in concentration estimates. The sample relative standard deviation is often used to describe the precision of a set of replicate measurements. For example, this statistic is used in the initial precision and recovery experiment validation criteria of Method 1623 (USEPA, 2005b). The population relative standard deviation is σ/μ × 100%. The relative standard deviation (RSD) of the unbiased concentration estimate can be calculated by dividing the square root of Equation 8.14 by concentration to obtain Equation 8.16 (Schmidt *et al.*, 2008). One advantage of using Equation 8.16 (rather than Equation 8.14) is that the precision of concentration estimates (as a function of μ_p and σ_p^2) can be investigated as a function of the dimensionless value cV rather than having to specify a concentration. This product is the expected number of particles contained in a sample of volume V drawn from a source with concentration c, so precision can be addressed in terms of numbers of microorganisms per sample.

$$RSD\left[\frac{x}{V\mu_p}\right] = \sqrt{\frac{1}{cV\mu_p} + \frac{\sigma_p^2}{\mu_p^2}} \times 100\% \qquad (8.16)$$

Figure 8.4 shows the precision of concentration estimates as a function of the expected number of particles per sample using two analytical recovery scenarios: $\mu_p = 0.3$ and $\sigma_p^2 = 0.007$ in Scenario 1, and $\mu_p = 0.8$ and $\sigma_p^2 = 0.002$ in Scenario 2. It shows that the relative standard deviation is substantially reduced by increasing the expected number of particles per sample beyond 10 (Schmidt *et al.*, 2008), but that there is little additional value in collecting samples with more than 500 microorganisms. The minimum relative standard deviation is shown to increase as σ_p^2/μ_p^2 increases. The figure also shows the over-dispersion of counts (with respect to the Poisson distribution) that is due to variability in analytical recovery (indicated by the separation between the SAE and RSD curves). The curves associated with the first scenario show over-dispersion whenever the expected number of microorganisms in the sample is greater than about 5. In the second scenario, the counts are not substantially over-dispersed because the variability in analytical recovery is low. Most importantly, this analysis shows that it behoves analysts to choose sample volumes (wherever possible) that will result in samples containing an average of 10 or more analyte particles. Samples that typically contain fewer than 10 analyte particles will yield relatively imprecise concentration estimates (unless several replicates are averaged). This demonstrates that quantitative analysis approaches that address uncertainty in unbiased concentration estimates are particularly essential when counts below 10 are obtained.

Effect of replication upon concentration estimates

Replication will enable collection of more precise concentration estimates (if the replicates are suitably pooled to yield a single unbiased concentration estimate) because a larger volume of the source will have been enumerated than what is enumerated in any one of the samples. Additionally the mean recovery of several samples will be closer to the mean recovery of the enumeration method. In effect, pooling replicates averages out measurement error (if the concentration estimates are unbiased).

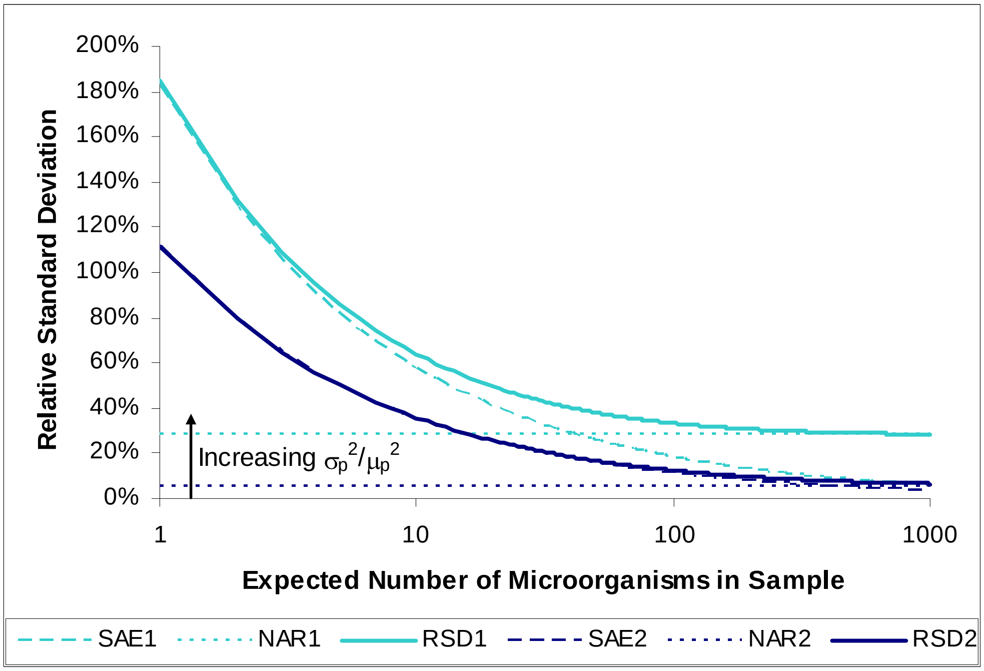


Figure 8.4: Relative standard deviation of concentration estimates
(after Schmidt *et al.*, 2008)

'SAE' represents the minimum relative standard deviation due to seeding and analytical error if σ_p^2 were zero, 'NAR' represents the minimum relative standard deviation as $cV\mu_p$ goes to infinity, and 'RSD' represents the relative standard deviation of the concentration estimates. Scenario 1 uses $\mu_p = 0.3$, $\sigma_p^2 = 0.007$, and Scenario 2 uses $\mu_p = 0.8$, $\sigma_p^2 = 0.002$.

To investigate the value of replication without the obvious effect that averaging over larger cumulative sample volumes yields a more precise concentration estimate, this section addresses the situation in which a total volume of 100 L is enumerated in a number of equal-volume replicates. The resulting concentration estimate is the sample mean of replicate concentration estimates. The same approach was used in Section 5.6.3. If the counts (and associated concentration estimates) are approximately normally distributed or the number of replicates (r) is large, then $\bar{c} \sim NORMAL(\mu,\sigma^2/r)$ according to the central limit theorem. The population mean of the concentration estimates is $\mu = E[x/V\mu_p] = c$ and the population variance is $\sigma^2 = Var[x/V\mu_p]$ (from Equation 8.14). Accordingly, if the normality assumption holds, a 95% probability interval for $\bar{c}$ can be calculated using Equation 8.17 (Schmidt *et al.*, 2008); otherwise, probability intervals could be evaluated by Monte Carlo. Equation 8.18 calculates the probability of a sample mean concentration that is below 0 (if normality is assumed) and can be used as a quick diagnostic for non-normality (*e.g.* if $\alpha > 0.01$). In Section 5.5.1, it was shown that departures of count distributions from normality are greatest when the counts are low. Subdividing the total volume into replicates will reduce the number of microorganisms in each sample (and increase departure of the counts from normality), but the increase in replication will simultaneously make the distribution of sample means tend towards normality.

$$c - \frac{z_{0.025}}{\sqrt{r}} \cdot \sqrt{\frac{c}{V\mu_p} + \frac{c^2\sigma_p^2}{\mu_p^2}} < \bar{c} < c + \frac{z_{0.025}}{\sqrt{r}} \cdot \sqrt{\frac{c}{V\mu_p} + \frac{c^2\sigma_p^2}{\mu_p^2}} \tag{8.17}$$

$$\alpha = P\left(Z < -\left(\frac{1}{rcV\mu_p} + \frac{\sigma_p^2}{r\mu_p^2} \right)^{-0.5} \right) \tag{8.18}$$

To illustrate use of this statistical tool, Equation 8.17 was used with two analytical recovery scenarios: Scenario 1 is a situation in which analytical recovery is low and highly variable (μ_p = 0.3, σ_p^2 = 0.007), and Scenario 2 is a situation in which analytical recovery is high with low variability (μ_p = 0.8, σ_p^2 = 0.002). The example uses a concentration of 100 microorganisms/L and a total volume of 100 L and considers alternatives from a single 100 L sample through to twenty 5 L samples. The highest value of α in any of the presented analyses is 0.000173 (which does not indicate non-normality). This is likely because the expected total number of microorganisms in the 100 L total volume is 10,000. The resulting 95% probability intervals for $\bar{c}$ are plotted in Figure 8.5. This figure shows that replication is very beneficial in Scenario 1 (which has highly variable analytical recovery), but not especially worthwhile in Scenario 2 (which has relatively constant analytical recovery). This is consistent with the findings in Section 5.6.3. Variability in analytical recovery causes over-dispersion of counts with respect to the Poisson distribution, which causes replication to be beneficial. In Scenario 2, the lower variability in analytical recovery reduces this over-dispersion, and replication is not worthwhile in the case of Poisson-distributed counts (Haas, 1993). As described previously, variance decomposition can be used to evaluate over-dispersion that is due to non-constant analytical recovery (*e.g.* the second term in Equation 8.14). Under- or over-dispersion with respect to the Poisson distribution that is due to non-random particle dispersion in the source (as opposed to variability in analytical recovery, concentration, or sample volume) is addressed in Appendix E.2.3. Variance decomposition is a convenient method to investigate the impacts of several factors upon the precision of concentration estimates and to evaluate experimental design.

8.2.3 Variance Decomposition with Variable Concentration

This section further applies variance decomposition to account for the situation in which concentration is a random variable (*e.g.* when there is temporal concentration variability) with mean μ_c and variance σ_c^2. In this case, $E[x/V\mu_p] = \mu_c$, so the count divided by the mean analytical recovery of the enumeration method is an unbiased estimate of the mean concentration. The associated variance is shown in Equation 8.19, and is derived in Appendix E. This equation clearly shows that the variance of a set of concentration estimates is always a biased estimate of the variability in concentration, although this bias diminishes as the first two terms approach zero. Like variability in the analytical recovery of the enumeration method, variability in concentration also yields counts that are over-dispersed with respect to the Poisson distribution. Variability in sample volumes would have a similar effect, but this variability is not addressed herein.

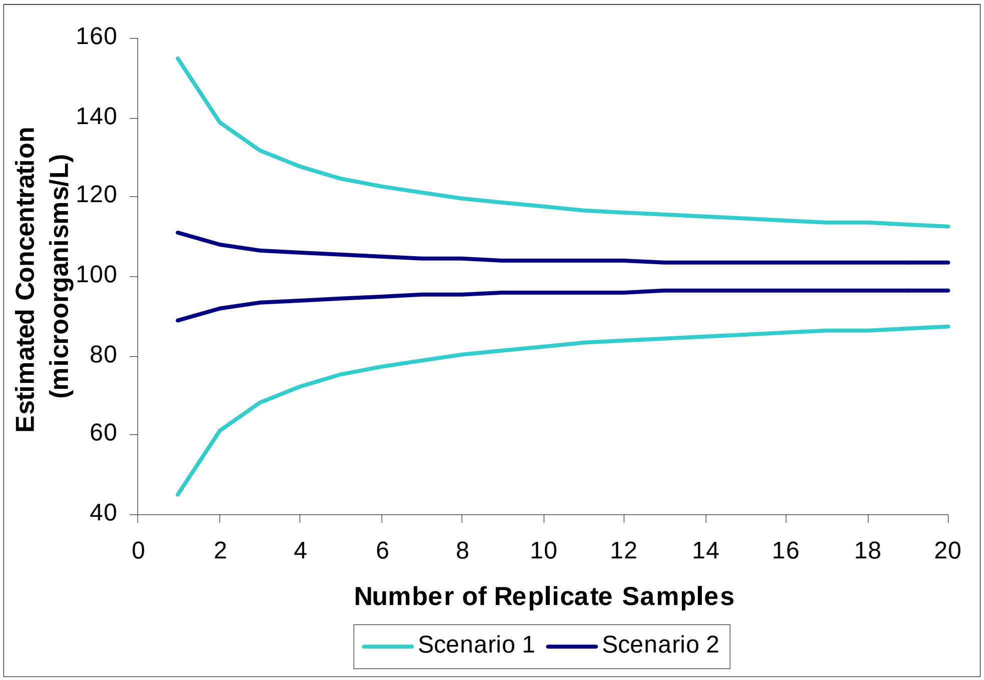


Figure 8.5: Effect of replication upon concentration estimates
(after Schmidt *et al.*, 2008)

The concentration is 100 microorganisms/L and the total volume of the equal-volume replicates is 100 L. Scenario 1 uses $\mu_p = 0.3$, $\sigma_p^2 = 0.007$, and Scenario 2 uses $\mu_p = 0.8$, $\sigma_p^2 = 0.002$.

$$Var\left[\frac{x}{V\mu_p}\right] = \frac{\mu_c}{V\mu_p} + \frac{\left(\mu_c^2 + \sigma_c^2\right)\sigma_p^2}{\mu_p^2} + \sigma_c^2 \qquad (8.19)$$

Like Equation 8.14, Equation 8.19 could be used as an experimental design tool. This is not addressed herein because many of the results would be comparable to Section 8.2.2 and because this topic was addressed using Monte Carlo simulations in Section 6.2. The precision of the sample mean or sample variance of a set of independent unbiased concentration estimates could be addressed using the central limit theorem and the chi-squared distribution (in a similar fashion to Section 8.14) if the concentration estimates were assumed to be approximately normally distributed. A simple diagnostic test for non-normality that is similar to Equation 8.18 could be constructed using the population variance calculated in Equation 8.19. Such an analysis could be further expanded (here and with the recovery estimates in Section 8.1) to compare probability intervals that ignore measurement errors to those that address them. This would further prove the findings in Sections 4.4.5 and 5.5.1 that the power of hypothesis tests to classify a departure of the sample mean from the null hypothesis as significant is reduced by measurement error. Accordingly, further investigations using variance decomposition are recommended.

8.3 PRACTICAL IMPLICATIONS

The findings presented in this chapter show that variance decomposition, coupled with simple probabilistic models is a powerful and relatively simple statistical tool to investigate many issues associated with measurement errors in analytical recovery and concentration estimates. In particular, it was used to demonstrate and quantify various types of bias, and as an experimental design tool. A summary of specific conclusions that were supported by this work is provided below.

Quantification of Analytical Recovery

- The count obtained from a seeded sample divided by the presumed number of seeded particles is an unbiased estimate of analytical recovery, but the variance of such estimates is an over-estimate of the variability in analytical recovery.
- Increasing the quantity of seeded particles can reduce measurement errors in recovery estimates (if feasible and if it does not adversely affect analytical recovery and the representativeness of the experiment), but there is a threshold beyond which further increases are not beneficial; this threshold occurs at a lower quantity of particles when analytical recovery is highly variable, and equations are provided herein to calculate desirable quantities of seeded particles.
- It is preferable to seed recovery samples with precisely known numbers of analyte particles rather than with a specific volume from a stock of known concentration, and the difference in measurement error between the two methods is greatest when mean analytical recovery is high.
- Seeding a recovery sample with precisely known numbers of particles and analyzing only a fraction of the well-mixed sample can result in unbiased recovery estimates that are greater than 100% and at least partially negates the benefits of seeding with precisely known numbers of analyte particles (as opposed to an aliquot of stock suspension).
- Increasing the number of replicate samples in a recovery experiment increases the precision of the sample mean and sample standard deviation of the recovery estimates, and a statistical tool is presented herein to provide a graphical and quantitative measure of this precision.

Quantification of Concentration (and the variability therein)

- The count obtained from a sample divided by the enumerated sample volume is a biased estimate of concentration unless the recovery is 100%; analysts must divide the count per unit volume analyzed by the mean analytical recovery of the enumeration method or a sample-specific recovery estimate in order to report unbiased concentration estimates.
- Counts of analyte particles can be over-dispersed with respect to the Poisson distribution due to variability in sample volume, concentration, or analytical recovery of the enumeration method; non-random dispersion of particles in the source cannot be inferred

from over-dispersed counts unless the counts are replicates and the variability in analytical recovery of the enumeration method has been addressed.

- The sample variance of a set of unbiased concentration estimates is a biased estimate of the variance of concentration due to unavoidable measurement errors; measurement errors cause fully replicated concentration estimates to vary.
- Increasing the sample volume (where feasible) can often reduce the measurement error in a concentration estimate, but there is a threshold beyond which further increases are not beneficial; this threshold occurs at a lower volume when analytical recovery is highly variable, when the concentration is high, or the mean analytical recovery is low, and an equation is provided herein to calculate a desirable sample volume given a hypothetical value of concentration.
- The relative standard deviation of a set of unbiased concentration estimates can be substantially reduced by ensuring that sample volumes are large enough to contain an average of at least 10 particles; samples that typically contain fewer than ten particles will yield highly uncertain unbiased concentration estimates, which especially necessitates statistical analysis methods that address uncertainty due to measurement error.
- Improving the analytical recovery of an enumeration method (*i.e.* increasing the sample mean or decreasing the variability in analytical recovery) may not have a substantial impact upon the precision of concentration estimates; an equation is provided herein that can be used to quantify the effect of changes in analytical recovery upon the precision of concentration estimates.
- When the concentration is constant, replication (as opposed to enumerating a single sample with the same total volume) is only beneficial if the variability in analytical recovery is high; a statistical tool is presented herein to provide a graphical and quantitative measure of increased precision due to replication.

Chapter 9
Conclusions and Future Research

Parameters associated with the detection and enumeration of microorganisms and discrete particles in water such as the analytical recovery of an enumeration method, the concentration of analyte particles in water, the log-reduction of a treatment process, and the sensitivity of a detection method cannot be measured exactly. There are unavoidable random errors that make estimates of these parameters imprecise and potentially inaccurate. This presents two problems: (1) how to analyze the data with appropriate regard for measurement error so that reliable scientific conclusions and management decisions can be made, and (2) how to design future experiments so that more informative (*i.e.* less imprecise) data can be obtained with the available resources. Herein, probabilistic modelling was used to address these issues in many scenarios that are of general interest in quantitative microbiology and that are of particular interest to the drinking water industry. The production of microbiologically safe drinking water depends upon reliable science (*e.g.* development of new methods to quantify microorganisms in water, source water monitoring, treatment technology research, risk analysis), which in turn depends upon collection of reliable data and use of appropriate data analysis strategies. This chapter provides a summary of the conclusions and implications of this research (Section 9.1) and discusses several areas in which further research is necessary (Section 9.2).

9.1 CONCLUSIONS AND IMPLICATIONS

This section provides a summary of the principal conclusions of this research (shown in bold font) and their respective implications. The specific objectives or sections of the thesis to which each conclusion or implication pertains are provided (where applicable).

1) Measurement errors must be addressed because most scientific research and management decisions need to be based upon actual values of the parameter being measured rather than the imprecise and potentially inaccurate measurements (or estimates) that are obtained. The reliability of any result or decision cannot be demonstrated without considering the inherent measurement error in the data because some data are more reliable than others and data with considerable measurement error obscure reality. Many parameters associated with the enumeration of microorganisms and discrete particles in water cannot be measured exactly.

- Measurement error must be addressed in the analysis of analytical recovery because the fraction of seeded microorganisms or discrete particles that are observed in a sample is an imprecise estimate of analytical recovery (Objective 1, Section 4.2).

- Measurement error must be addressed in the analysis of microorganism and discrete particle concentrations because the count per unit of volume analyzed is an imprecise and potentially inaccurate estimate of concentration (Objective 2, Section 5.1).
- Measurement error must be addressed in the analysis of the log-reduction of a treatment process because the logarithm of the ratio of pre- and post-treatment concentration estimates is an imprecise and potentially inaccurate estimate of log-reduction (Objective 2, Section 5.4).
- Measurement error must be addressed in the analysis of the sensitivity of a detection method because the fraction of replicate samples in which at least one microorganism or discrete particle is observed is an imprecise estimate of sensitivity (Objective 4, Section 7.2).
- The reliability of data cannot be ascertained when the method that is used to obtain them is described insufficiently or the raw data are not provided (Sections 4.4.1 and 5.4.4).

2) Direct analysis of data without consideration of measurement error typically results in bias. Without bias, a reported result is only uncertain due to unavoidable random error; with bias, the reported result errs systematically from the true value. Bias can lead to misguided conclusions or decisions in some cases, and to excessive conservatism in others.

- Recovery estimates are more variable than recovery itself unless the measurement error is trivially small. Analysis of recovery estimates without consideration of measurement errors will result in over-estimated variance and incorrectly fitted distributions that describe the variability in recovery (Objective 1, Sections 4.3 and 4.4).
- The count of microorganisms or discrete particles per unit of volume analyzed is a biased estimate of concentration unless the analytical recovery is 100%. The count per unit volume must be divided by the mean analytical recovery of the enumeration method or a sample-specific recovery estimate to calibrate the concentration estimate to the actual concentration, but these unbiased estimates are still uncertain (Sections 5.6, 6.2, and 8.2).
- The logarithm of the ratio of pre- and post-treatment concentration estimates is a biased estimate of the log-reduction of a treatment process unless both concentration estimates are unbiased or both are equally biased (Section 5.4.4).
- The variability of concentration estimates (*e.g.* over time) is a biased estimate of the variability in concentration itself unless the measurement error is trivially small (Sections 6.2 and 8.2.3).
- t-tests on mean analytical recovery or mean concentration are biased (unless the measurements are precise) because unaddressed measurement error results in over-estimated variance and this reduces the power of the test to classify an observed departure from a false null hypothesis as statistically significant (Sections 4.4.5 and

5.5.1). Moreover, enumeration data are often not approximately normally distributed and the data may not have constant measurement errors, both of which compromise the validity of t-tests and many other conventional tests (*e.g.* chi-squared tests, F-tests, and ANOVA) unless the data are suitably transformed.

- The Long Term 2 Enhanced Surface Water Treatment Rule (USEPA, 2006) should be revised to address the analytical recovery associated with *Cryptosporidium* monitoring data because public water systems that achieve high analytical recovery are more likely to incur mandatory treatment enhancements than those that achieve low analytical recovery (Section 6.4).

3) Probabilistic models that address random errors in enumeration data with respect to the unknown parameters that are of interest to the analyst are the basis for appropriate statistical analysis of data (*i.e.* analysis that is not biased by ignoring errors in the data). These models should be used to obtain appropriate point estimates of unknown parameters, and Bayes' theorem can subsequently be used to quantify the uncertainty in all the unknown parameters. The likelihood function describes the relative plausibility of alternative values of the unknown parameter(s) given data and a probabilistic model that describes the joint probability of the data as a function of the parameters. Bayes' theorem expands upon this concept by incorporating potentially subjective prior information and by providing a quantitative description of the uncertainty in the unknown parameters in the form of a posterior distribution.

- Maximum likelihood estimation can be used to obtain point estimates of parameters for the distribution that describes random variability in analytical recovery with consideration of the measurement error in replicate recovery estimates (Objective 1, Section 4.4.1).
- Bayes' theorem and Gibbs sampling can be implemented to quantify the uncertainty in non-constant analytical recovery distribution parameters given replicate recovery data (Section 4.4.3).
- Bayes' theorem and either numerical integration or Gibbs sampling can be implemented to quantify the uncertainty in a concentration estimate given a single datum or replicate data (Objective 2, Section 5.3). This analysis approach also enables rigorous interpretation of non-detect samples (Objective 4, Sections 5.3.3 and 7.1.2).
- Bayes' theorem and Gibbs sampling can be implemented to quantify the uncertainty in a log-reduction estimate (Objective 2, Section 5.4.3).
- Bayes' theorem and Gibbs sampling can be implemented to quantify the uncertainty in parameters for a temporal concentration variability distribution by addressing the uncertainty in the individual concentration estimates (Section 6.3.1).
- Bayes' theorem and Gibbs sampling can be implemented to quantify the uncertainty in the sensitivity of a detection method given analytical recovery data (Objective 4, Section 7.2).

- Bayesian analysis of non-detect samples shows that reporting non-detect samples as a concentration of <1 microorganism per unit of volume analyzed (the method detection limit) is a misleading representation of uncertainty. The posterior probability that the concentration is greater than the method detection limit is demonstrated to be at least 36% (Objective 4, Section 7.1.2).

4) Monte Carlo Quantitative Microbial Risk Assessment (QMRA) for drinking water must address the measurement errors in pathogen concentration data using appropriate data analysis approaches in order to calculated risk estimates that are quantitatively predictive. Exposure assessment depends upon characterization of the pathogen concentration in the water and the variability therein, but neither can be quantified appropriately without considering the uncertainty in the individual concentration estimates (Objective 3, Section 6.5).

- Analytical recovery must not be ignored in QMRA unless it is consistently 100%. Assuming that incomplete analytical recovery and incomplete infectivity of the observed pathogens are counter-acting errors will usually lead to bias (Section 6.5.2).

- Measurement error in the concentration estimates must be addressed in Monte Carlo QMRA because unbiased concentration estimates will be more variable than the actual pathogen concentration (Section 8.2.3). Appropriate data analysis evaluates the portion of the variability among the concentration estimates that is due to variability in the concentration itself by considering the uncertainty in each concentration estimate.

- Microorganism counts (or counts per unit volume) and analytical recovery are generally not statistically independent. QMRA models that falsely assume this independence artificially inflate the variability in concentration rather than evaluating the portion of the existing variability that is actually due to variability in the pathogen concentration. Modelling analytical recovery as a dependent random variable may reduce this bias, but depends upon extensive case-specific data to evaluate correlation (Section 6.5.2).

- Manipulating non-detect data into positive concentration values leads to bias. All concentration estimates are uncertain due to unavoidable measurement error, and this uncertainty must be addressed for non-detects and positive counts alike (Section 6.5.1).

- Probabilistic models that address random variability in the pathogen concentration over time as well as random measurement errors in the concentration data can be used (with Bayes' theorem and Gibbs sampling) in Monte Carlo QMRA to evaluate the temporal concentration variability, and uncertainty therein, appropriately (Sections 6.3 and 6.5.3).

5) To make better use of available resources and ensure that acceptably precise data are obtained in future experiments and monitoring programs, probabilistic models should be used evaluate the anticipated measurement error and the case-specific factors that affect it most. There are often many alternative strategies to obtain more precise data, but some are more impactful than others. A probabilistic model that describes the random errors in enumeration data (and that is appropriate for the methodology that will be used to collect the data) can be used to evaluate the anticipated measurement error and to compare the alterative experimental designs (Objective 5).

- Increasing the number of seeded microorganisms or discrete particles (if it does not affect the representativeness of the recovery data) and improving the precision of the seed dose will result in more precise recovery estimates. The effect of the seed dose (and possible imprecision thereof) and the number of replicate samples in a recovery experiment upon the mean and standard deviation of the recovery estimates can be evaluated using probabilistic models and either Monte Carlo simulation or variance decomposition (Sections 4.3 and 8.1.4). Specific findings are summarized in Sections 4.5 and 8.3.
- Processing larger sample volumes, processing replicate samples, using internal standards to obtain sample-specific recovery estimates, or improving the enumeration methodology so that the mean analytical recovery is closer to 100% or the variability in analytical recovery is lower will result in more precise concentration estimates. Alternative strategies can be compared using Monte Carlo simulation or variance decomposition (Sections 6.2 and 8.2.2). Using credible intervals arising from Bayesian analysis of hypothetical data is a comparatively inefficient approach to evaluate the effects of various factors upon the uncertainty in concentration and log-reduction estimates (Section 5.6).
    - More precise concentration estimates can often be obtained by increasing the sample volume, especially so that samples will contain at least 10 microorganisms or discrete particles on average.
    - Replication, as opposed to processing a single sample with the same total volume, is most effective when analytical recovery varies substantially among replicate samples.
    - Using internal standards to obtain sample-specific recovery estimates when enumerating indigenous microorganisms is only beneficial when variability in analytical recovery contributes substantially to the variability in the data.
    - Using methods with improved analytical recovery does not always substantially improve the precision of concentration estimates (especially when the counts are low). Method development should focus upon enabling larger sample volumes to be processed or upon logistical concerns such as reducing cost and complexity, or improving efficiency.
- This type of analysis can also be used to determine if less rigorously designed experiments would yield similarly reliable data. For example, there is no need to process a 100 L sample if a 5 L sample would yield a comparably reliable concentration estimate.

6) The test (based on Fisher’s index of dispersion) that has often been used to evaluate possible clumping of microorganisms or particles in water is often inappropriate for this purpose. Most probabilistic models associated with the detection or enumeration of microorganisms in water assume that the number of microorganisms contained in a sample drawn from a homogeneous source is Poisson-distributed (which implicitly assumes that the microorganisms are randomly distributed in the source with constant concentration). Counts that have a variance that is substantially different from their mean are under- or over-dispersed with respect to the Poisson distribution, and this can be tested using Fisher’s index of dispersion (Section 3.2.1). Over-dispersion is often attributed to a non-random (*i.e.* clumped) particle distribution in the source.

- Significant over-dispersion does not conclusively indicate clumping of microorganisms or discrete particles in a source unless the samples are replicates (*i.e.* the concentration in the source is constant), the sample volumes are equal, and the analytical recovery of the enumeration method is constant among samples. Clumping should be evaluated by visual inspection where possible (Section 5.2.1).
- Insignificant under- or over-dispersion of counts does not confirm an approximately random particle distribution throughout the source at all possible sample volumes: the outcome applies only to the sample volume upon which the test is based (Section 5.2.2).

9.2 FUTURE RESEARCH TOPICS

This section provides a brief discussion of several areas in which the concepts presented herein can be expanded to further improve the statistical analysis of microorganism and discrete particle enumeration data. In addition to these future research topics, it is necessary to develop software that will make the statistical tools presented herein more available to practitioners.

Evaluation of the Effects of Seed Dose upon Analytical Recovery

The models that are presented herein assume that analytical recovery is independent of the quantity (or concentration) of seeded microorganisms or discrete particles (at least within some practically relevant range). This assumption needs to be evaluated using a variety of seed doses and/or sample volumes with some common enumeration methods such as Method 1623 (USEPA, 2005b). It has been demonstrated herein that lower seed doses and seeding error increase the variability of replicate recovery estimates (because the estimates are more imprecise) even if the distribution of actual analytical recovery values remains the same. Therefore, experiments conducted to evaluate the effect of seed dose upon analytical recovery must use statistical approaches that address the measurement errors in the recovery estimates. Gibbs sampling, using an appropriate probabilistic model for the recovery data, can be used to evaluate the difference in mean analytical recovery or the ratio of the variances of analytical recovery for two populations (representing different seed doses) as shown in Figure 4.26. Posterior predictive recovery distributions for various seed doses can also be compared to identify possible differences in the distribution of actual recovery values. Differences among these distributions, however, may also reflect varying levels of uncertainty due to the size of the respective datasets or the precision of individual recovery estimates. Further research is also needed to evaluate how the Bayesian quantitative analysis approaches proposed herein compare to other parametric and non-parametric hypothesis tests that have commonly been used in method development research to interpret recovery experiment data.

Evaluation of Temporal Log-reduction Variability

The log-reduction model developed herein quantifies the uncertainty in an individual log-reduction estimate by evaluating the uncertainty in the pre- and post-treatment concentration estimates upon which it is based. This model should be expanded to address temporal log-reduction variability because information about this variability is required in drinking water Monte Carlo Quantitative Microbial Risk Assessment. The temporal log-reduction variability model should be expanded from the existing probabilistic log-reduction models (in a similar fashion to the temporal concentration variability expansion of the models that are used to quantify uncertainty in individual concentration estimates). Additionally, further research is needed to evaluate the appropriateness of ‘probability of passage’ models that are used to evaluate log-reduction and the temporal variability therein. At a minimum, such models need to be expanded to address the analytical recovery of the enumeration method.

Incorporation of the Temporal Concentration Variability Model that Addresses Measurement Errors into Monte Carlo Quantitative Microbial Risk Assessment

Quantitatively predictive QMRA must appropriately evaluate the pathogen concentration in the source and the temporal variability therein. At present, the literature addressing Monte Carlo QMRA for *Cryptosporidium* in drinking water shows that measurement error in concentration estimates is often ignored or handled incorrectly (especially when non-detect samples are manipulated into non-zero values) and that analytical recovery (if it is addressed at all) is often incorporated into the model in ways that can result in over-predicted risks. The modelling approach used herein incorporates information about analytical recovery into the evaluation of uncertainty in individual concentration estimates and this uncertainty is subsequently considered when evaluating the temporal concentration variability (and uncertainty therein). It has been demonstrated that this approach can lead to a distribution describing temporal concentration variability that is very different from the results of conventional analyses and it has been hypothesized that this may, in some cases, have a substantial impact upon computed risks. Further research is needed to evaluate whether or not more appropriate strategies to model temporal concentration variability (and to incorporate information about analytical recovery into the model) have a substantial impact upon computed risks. Further research is also needed to develop appropriate strategies to incorporate information about the infectivity of the observed pathogens (and uncertainty therein) and the log-reduction of the treatment process (and the variability and uncertainty therein) into exposure assessment. Probabilistic modelling research is also required to address the random errors in dose response experiments and to quantify the uncertainty in dose-response models appropriately.

Use of Other Markov Chain Monte Carlo Algorithms

Gibbs sampling, a form of Markov Chain Monte Carlo, was used herein to evaluate complicated posterior distributions resulting from the application of Bayes' theorem to various probabilistic models. Uniform priors (or other simple relatively uninformative priors) were used in most of these Bayesian models, so Gibbs sampling was relatively easy to implement. Nonetheless, some of the models resulted in conditional posterior distributions from which generation of random deviates was complicated. If more complicated priors are used, then Gibbs sampling may become more computationally cumbersome. Implementation of the Metropolis-Hastings algorithm (Gelman *et al.*, 2004) will resolve the issue of complicated conditional posterior distributions. This algorithm, however, can lead to worse mixing of the Markov Chains and generally requires more mathematical expertise than Gibbs sampling to implement it correctly.

Modelling Measurement Errors in Other Types of Methods Used to Quantify Microorganisms

The models and associated statistical analysis tools presented herein address random errors in the collection of various types of enumeration-based data. Many of the concepts apply to all methods that are used to quantify microorganisms in water (and possibly also in other media), but further research is required to develop models that represent measurement errors in other types of methods. Further research is required for plating methods in which multiple counts are obtained from serial dilutions because the counts would not be independent and the error may vary depending on the number of dilutions. Probabilistic models are already used in most probable number methods, but these models assume Poisson-distributed random sampling error and essentially assume 100% analytical recovery. Finally, probabilistic modelling is needed for quantitative polymerase chain reaction methods because there is random sampling error in sample collection, analytical error in nucleic acid extraction, and further analytical error in the polymerase chain reaction.

Appendix A
Algebraic Proofs and Optimization of Likelihood Functions

This appendix contains proof that two nested binomial distributions can be combined into a single binomial distribution (Appendix A.1), discussion of the optimization of various likelihood functions (Appendix A.2), and proof that the conditional posterior distributions for which acceptance-rejection sampling is used are unimodal (Appendix A.3).

A.1 Combining Nested Binomial Distributions

In Section 4.2.5, the analytical errors associated with partial sample analysis (with fraction θ) and imperfect analytical recovery (p) were modelled with nested binomial distributions. These were consolidated into a single binomial model with the probability of each particle being observed equal to θp. This was summarized by Equation 4.9 (which is also shown below), and the proof follows.

$$\sum_{m=x}^{n}\left[\binom{n}{m}\theta^{m}(1-\theta)^{n-m}\right]\left[\binom{m}{x}p^{x}(1-p)^{m-x}\right]=\binom{n}{x}(\theta p)^{x}(1-\theta p)^{n-x}$$

Proof:

$$\begin{aligned}
\text{LHS} &= \sum_{m=x}^{n}\left[\binom{n}{m}\theta^{m}(1-\theta)^{n-m}\right]\left[\binom{m}{x}p^{x}(1-\theta)^{m-x}\right] \\
&= \frac{n!}{x!}p^{x}\cdot\sum_{m=x}^{n}\left[\frac{1}{(n-m)!(m-x)!}\theta^{m}(1-\theta)^{n-m}(1-p)^{m-x}\right] \\
&= \frac{n!}{x!(n-x)!}(\theta p)^{x}\cdot\sum_{m=x}^{n}\left[\frac{(n-x)!}{(n-m)!(m-x)!}\theta^{m-x}(1-\theta)^{n-m}(1-p)^{m-x}\right] \\
&= \binom{n}{x}(\theta p)^{x}\cdot\sum_{I=0}^{n-x}\left[\binom{n-x}{I}(\theta-\theta p)^{I}(1-\theta)^{n-x-I}\right] && \text{Sub } I = m - x \\
&= \binom{n}{x}(\theta p)^{x}(1-\theta p)^{n-x}\cdot\sum_{I=0}^{n-x}\left[\binom{n-x}{I}\left(\frac{\theta-\theta p}{1-\theta p}\right)^{I}\left(\frac{1-\theta}{1-\theta p}\right)^{n-x-I}\right] && \text{Note } (1-\theta)+(\theta-\theta p)=1-\theta p \\
&= \binom{n}{x}(\theta p)^{x}(1-\theta p)^{n-x}\sum_{I=0}^{n-x}\left[\binom{n-x}{I}\left(\frac{\theta-\theta p}{1-\theta p}\right)^{I}\left(1-\frac{\theta-\theta p}{1-\theta p}\right)^{n-x-I}\right] && \text{A binomial summation!} \\
&= \binom{n}{x}(\theta p)^{x}(1-\theta p)^{n-x} \\
&= \text{RHS}
\end{aligned}$$

A.2 OPTIMIZATION OF LIKELIHOOD FUNCTIONS

This section describes optimization of the likelihood functions discussed in Section 4.4.1. The models with non-constant analytical recovery each have two unknown parameters; it is difficult to find the critical points because it would be necessary to find points at which the directional derivative is zero in all directions or in which the directional derivative is infinite in at least one direction. These particular functions are also difficult to maximize because the derivatives of functions that include the gamma function *Γ*() cannot be evaluated explicitly. The 'binomial-only' and 'Poisson-only' models, on the other hand, are easy to optimize because they each have only one unknown parameter and because the derivative can be evaluated explicitly. Accordingly, an explicit formula for the maximum of these functions can be derived quite easily. For the two-parameter models, first and second derivatives for each parameter are evaluated to describe the shape of the likelihood functions. The notation *G*(.) is used for the natural logarithm of the gamma function *Γ*(.). Some useful properties of the gamma function are described below.

- $G'(x+y)-G'(x)>0$ for all positive *y* because $G''(x)>0$ for all *x* > 0 (*i.e.* $G'(x)$ is monotonic increasing).
- $G''(x+y)-G''(x)<0$ for all positive *y* because $G^{(3)}(x)<0$ for all *x* > 0 (*i.e.* $G''(x)$ is monotonic decreasing).

A.2.1 'Beta-only' Recovery Model

Likelihood

$$L(a,b)=\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)^{r}\left(\prod_{i=1}^{r}\hat{p}_i\right)^{a-1}\left(\prod_{i=1}^{r}(1-\hat{p}_i)\right)^{b-1} \qquad a>0,\ b>0$$

log-Likelihood

$$l(a,b)=r\big(G(a+b)-G(a)-G(b)\big)+(a-1)\cdot\sum_{i=1}^{r}\ln(\hat{p}_i)+(b-1)\cdot\sum_{i=1}^{r}\ln(1-\hat{p}_i)$$

Partial Derivatives

$$\frac{\partial l}{\partial a}=r\big(G'(a+b)-G'(a)\big)+\sum_{i=1}^{r}\ln(\hat{p}_i) \qquad (1) \qquad \frac{\partial l}{\partial b}=r\big(G'(a+b)-G'(b)\big)+\sum_{i=1}^{r}\ln(1-\hat{p}_i) \qquad (2)$$

$$\frac{\partial^2 l}{\partial a^2}=r\big(G''(a+b)-G''(a)\big) \qquad (3) \qquad \frac{\partial^2 l}{\partial b^2}=r\big(G''(a+b)-G''(b)\big) \qquad (4)$$

Discussion

Equations 3 and 4 are negative for all values of *a* and *b*, so the log-likelihood function is concave down (*i.e.* it has a maximum) for either parameter when the other is fixed. By setting Equations 1 and 2 equal to zero, it is found that both parameters are independently optimized when

$G'(a)-G'(b)=\frac{1}{r}\sum_{i=1}^{r}\ln\left(\frac{1}{\hat{p}_i}-1\right)$. A critical point may exist along this line, but cannot be found without evaluating directional derivatives (which are not addressed herein). On the basis of evaluated likelihood surfaces, it is presumed that this log-likelihood function is concave down throughout the parameter space and that it has a unique maximum.

A.2.2 'Gamma-only' Recovery Model

Likelihood

$$L(\alpha,\beta)=\left(\frac{1}{\beta^{\alpha}\Gamma(\alpha)}\right)^{r}\left(\prod_{i=1}^{r}\hat{p}_i\right)^{\alpha-1}e^{-\sum_{i=1}^{r}\hat{p}_i/\beta} \qquad \alpha>0,\ \beta>0$$

log-Likelihood

$$l(\alpha,\beta)=-r\alpha\cdot\ln(\beta)-r\cdot G(\alpha)+(\alpha-1)\cdot\sum_{i=1}^{r}\ln(\hat{p}_i)-\frac{1}{\beta}\cdot\sum_{i=1}^{r}\hat{p}_i$$

Partial Derivatives

$$\frac{\partial l}{\partial\alpha}=-r\cdot\ln(\beta)-r\cdot G'(\alpha)+\sum_{i=1}^{r}\ln(\hat{p}_i) \qquad (1) \qquad \frac{\partial l}{\partial\beta}=-\frac{r\alpha}{\beta}+\frac{1}{\beta^2}\cdot\sum_{i=1}^{r}\hat{p}_i \qquad (2)$$

$$\frac{\partial^2 l}{\partial\alpha^2}=-r\cdot G''(\alpha) \qquad (3) \qquad \frac{\partial^2 l}{\partial\beta^2}=\frac{r\alpha}{\beta^2}-\frac{2}{\beta^3}\cdot\sum_{i=1}^{r}\hat{p}_i \qquad (4)$$

Discussion

Equation 3 is negative for all values of α, so this likelihood function has a maximum for any specified value of β. Equation 4 is negative if $\Sigma p_i < 0.5r\alpha\beta$; however, the expected value of Σp_i is $r\alpha\beta$ (because the expected value of p_i is $\alpha\beta$), so this second derivative is unlikely to be negative. Accordingly, this likelihood function may not have a maximum for any specified value of α. By setting Equation 2 equal to zero, it is found that a critical point for β will occur along the line $\alpha\beta=\frac{1}{r}\sum_{i=1}^{r}\hat{p}_i$ (*i.e.* the most likely value of the population mean is the sample mean). A critical point may exist along this line, but cannot be found without evaluating directional derivatives (which are not addressed herein). On the basis of evaluated likelihood surfaces, it is presumed that this likelihood function is concave down throughout the parameter space and that it has a unique maximum.

A.2.3 Beta-binomial Recovery Model

Likelihood

$$L(a,b) \propto \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \right)^r \prod_{i=1}^{r} \frac{\Gamma(x_i + a)\Gamma(n_i - x_i + b)}{\Gamma(n_i + a + b)} \qquad a > 0, b > 0$$

log-Likelihood

$$l(a,b) \propto r(G(a+b) - G(a) - G(b)) + \sum_{i=1}^{r} G(x_i + a) + \sum_{i=1}^{r} G(n_i - x_i + b) - \sum_{i=1}^{r} G(n_i + a + b)$$

Partial Derivatives

$$\frac{\partial l}{\partial a} = r(G'(a+b) - G'(a)) + \sum_{i=1}^{r} G'(x_i + a) - \sum_{i=1}^{r} G'(n_i + a + b) \tag{1}$$

$$\frac{\partial l}{\partial b} = r(G'(a+b) - G'(b)) + \sum_{i=1}^{r} G'(n_i - x_i + b) - \sum_{i=1}^{r} G'(n_i + a + b) \tag{2}$$

$$\frac{\partial^2 l}{\partial a^2} = r(G''(a+b) - G''(a)) + \sum_{i=1}^{r} G''(x_i + a) - \sum_{i=1}^{r} G''(n_i + a + b) \tag{3}$$

$$\frac{\partial^2 l}{\partial b^2} = r(G''(a+b) - G''(b)) + \sum_{i=1}^{r} G''(n_i - x_i + b) - \sum_{i=1}^{r} G''(n_i + a + b) \tag{4}$$

Discussion

Equation 3 is negative for all values of *a* and *b* because $\sum G''(x_i + a) - r \cdot G''(a) < 0$ and $\sum G''(n_i + a + b) - r \cdot G''(a+b) < 0$. Likewise, Equation 4 is negative because $\sum G''(n_i - x_i + b) - r \cdot G''(b) < 0$ and $\sum G''(n_i + a + b) - r \cdot G''(a+b) < 0$. Therefore, there the log-likelihood function is concave down (*i.e.* it has a maximum) for either parameter when the other is fixed. By setting Equations 1 and 2 equal to zero, it is found that both parameters are independently optimized when $\sum_{i=1}^{r} (G'(x_i + a) - G'(n_i + a + b) - G'(n_i - x_i + b) + G'(n_i + a + b)) = 0$. A critical point may exist along this line, but cannot be found without evaluating directional derivatives (which are not addressed herein). On the basis of evaluated likelihood surfaces, it is presumed that the beta-binomial log-likelihood function is concave down throughout the parameter space and that it has a unique maximum.

A.2.4 Beta-Poisson Recovery Model

Likelihood

$$L(a,b) \propto \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \right)^{r} \prod_{i=1}^{r} \int_{0}^{1} e^{-\lambda_i p} p^{x_i+a-1} (1-p)^{b-1} dp \qquad a > 0, b > 0$$

Let $I(a,b) = \int_{0}^{1} e^{-\lambda_i p} p^{x_i+a-1} (1-p)^{b-1} dp$

log-Likelihood

$$l(a,b) \propto r(G(a+b) - G(a) - G(b)) + \sum_{i=1}^{r} \ln(I(a,b))$$

Partial Derivatives

$$\frac{\partial l}{\partial a} = r(G'(a+b) - G'(a)) + \sum_{i=1}^{r} \frac{I'(a,b)}{I(a,b)} \qquad (1) \qquad \frac{\partial l}{\partial b} = r(G'(a+b) - G'(b)) + \sum_{i=1}^{r} \frac{I'(a,b)}{I(a,b)} \qquad (2)$$

$$\frac{\partial^2 l}{\partial a^2} = r(G''(a+b) - G''(a)) + \sum_{i=1}^{r} \frac{I''(a,b) \cdot I(a,b) - (I'(a,b))^2}{(I(a,b))^2} \qquad (3)$$

$$\frac{\partial^2 l}{\partial b^2} = r(G''(a+b) - G''(b)) + \sum_{i=1}^{r} \frac{I''(a,b) \cdot I(a,b) - (I'(a,b))^2}{(I(a,b))^2} \qquad (4)$$

Discussion

It is difficult to assess the shape of this log-likelihood function because of the integral *I*(*a*,*b*). On the basis of evaluated likelihood surfaces, it is presumed that the beta-binomial log-likelihood function is concave down throughout the parameter space and that it has a unique maximum.

A.2.5 Negative Binomial Recovery Model

Likelihood

$$L(\alpha,\beta) \propto \frac{1}{(\Gamma(\alpha))^{r}} \prod_{i=1}^{r} \frac{\Gamma(x_i+\alpha)\beta^{x_i}}{(\lambda_i\beta+1)^{x_i+\alpha}} \qquad \alpha > 0, \beta > 0$$

log-Likelihood

$$l(\alpha,\beta) \propto -r \cdot G(\alpha) + \beta^{\sum x_i} + \sum_{i=1}^{r} G(x_i+\alpha) - \sum_{i=1}^{r} (x_i \cdot \ln(\lambda_i\beta+1)) - \alpha \cdot \sum_{i=1}^{r} \ln(\lambda_i\beta+1)$$

Partial Derivatives

$$\frac{\partial l}{\partial \alpha} = -r \cdot G'(\alpha) + \sum_{i=1}^{r} G'(x_i+\alpha) - \sum_{i=1}^{r} \ln(\lambda_i\beta+1) \qquad (1)$$

$$\frac{\partial l}{\partial \beta} = \beta^{\left(\sum x_i\right)-1} - \sum_{i=1}^{r} \frac{x_i \lambda_i}{\lambda_i \beta + 1} - \alpha \cdot \sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i \beta + 1} \tag{2}$$

$$\frac{\partial^2 l}{\partial \alpha^2} = -r \cdot G''(\alpha) + \sum_{i=1}^{r} G''(x_i + \alpha) \tag{3}$$

$$\frac{\partial^2 l}{\partial \beta^2} = \beta^{\left(\sum x_i\right)-2} + \sum_{i=1}^{r} \frac{x_i \lambda_i^2}{(\lambda_i \beta + 1)^2} + \alpha \cdot \sum_{i=1}^{r} \frac{\lambda_i^2}{(\lambda_i \beta + 1)^2} \tag{4}$$

Discussion

Equation 3 is negative for all values of α because $G''(x_i + \alpha) - G''(\alpha) < 0$ for all α, so this likelihood function has a maximum for any specified value of β. Equation 4 could be positive or negative, so this likelihood function may not have a maximum for any specified value of α. On the basis of evaluated likelihood surfaces, it is presumed that this likelihood function is concave down throughout the parameter space and that it has a unique maximum.

A.2.6 'Binomial-only' Recovery Model

Likelihood

$$L(p) \propto (p)^{\sum x_i} (1-p)^{\sum (n_i - x_i)} \qquad 0 < p < 1$$

log-Likelihood

$$l(p) \propto \left(\sum_{i=1}^{r} x_i\right) \cdot \ln(p) + \left(\sum_{i=1}^{r} (n_i - x_i)\right) \cdot \ln(1-p)$$

Derivatives

$$\frac{dl}{dp} = \left(\sum_{i=1}^{r} x_i\right) \cdot \frac{1}{p} - \left(\sum_{i=1}^{r} (n_i - x_i)\right) \cdot \frac{1}{1-p} \tag{1}$$

$$\frac{d^2 l}{dp^2} = -\left(\sum_{i=1}^{r} x_i\right) \cdot \frac{1}{p^2} - \left(\sum_{i=1}^{r} (n_i - x_i)\right) \cdot \frac{1}{(1-p)^2} \tag{2}$$

Critical Point

$$0 = \left(\sum_{i=1}^{r} x_i\right) \cdot \frac{1}{p} - \left(\sum_{i=1}^{r} (n_i - x_i)\right) \cdot \frac{1}{1-p}$$

$$0 = \left(\sum_{i=1}^{r} x_i\right) - p \cdot \left(\sum_{i=1}^{r} x_i\right) - p \cdot \left(\sum_{i=1}^{r} n_i\right) + p \cdot \left(\sum_{i=1}^{r} x_i\right)$$

$$p = \frac{\sum x_i}{\sum n_i}$$

A.2.7 'Poisson-only' Recovery Model

Likelihood

$$L(p) \propto e^{-p\sum \lambda_i} p^{\sum x_i} \qquad p > 0$$

log-Likelihood

$$l(p) \propto -p \cdot \sum_{i=1}^{r} \lambda_i + \left(\sum_{i=1}^{r} x_i \right) \cdot \ln(p)$$

Derivatives

$$\frac{\partial l}{\partial p} = -\sum_{i=1}^{r} \lambda_i + \left(\sum_{i=1}^{r} x_i \right) \cdot \frac{1}{p} \qquad (1) \qquad \frac{\partial^2 l}{\partial p^2} = -\left(\sum_{i=1}^{r} x_i \right) \cdot \frac{1}{p^2} \qquad (2)$$

Critical Point

$$0 = -\sum_{i=1}^{r} \lambda_i + \left(\sum_{i=1}^{r} x_i \right) \cdot \frac{1}{p}$$

$$0 = -p \cdot \left(\sum_{i=1}^{r} \lambda_i \right) + \left(\sum_{i=1}^{r} x_i \right)$$

$$p = \frac{\sum x_i}{\sum \lambda_i}$$

A.3 Optimization of Conditional Posterior Distributions

There are several situations in this thesis in which a conditional posterior distribution used in Gibbs sampling cannot be arranged in the form of a distribution for which random number generating algorithms are readily available. Acceptance-rejection sampling is used in these cases, and the method that is used requires optimization of the conditional posterior. The two conditional posteriors that are of interest are shown below: the first is Equation 4.33, and the second is Equation 6.6.

$$f(\theta) \propto g(\theta) \cdot \left(\frac{\Gamma(\theta + s)}{\Gamma(\theta)} \right)^r t^\theta \qquad \theta > 0$$

$$Dfc(\rho) \propto g(\rho \mid \lambda) \cdot (\Gamma(\rho))^{-r} \left(\frac{1}{\lambda} \prod_{i=1}^{r} c_i \right)^\rho \qquad \rho > 0$$

A greedy search algorithm is used to optimize both types of conditional posterior distributions. This type of algorithm can converge upon a local maximum, so it is important to show that these functions are unimodal on the parameter space to ensure that the algorithm converges upon the global maximum.

Using the notation *G*(.) to represent the natural logarithm of the gamma function *Γ*(.), the shape of the first can be evaluated as follows.

$$\ln\left(f(\theta)\right)=\ln\left(g(\theta)\right)+r\cdot G(\theta+s)-r\cdot G(\theta)+\theta\cdot\ln(t)$$

$$\frac{d}{d\theta}\ln\left(f(\theta)\right)=\frac{d}{d\theta}\ln\left(g(\theta)\right)+r\cdot G'(\theta+s)-r\cdot G'(\theta)+\ln(t)$$

$$\frac{d^2}{d\theta^2}\ln\left(f(\theta)\right)=\frac{d^2}{d\theta^2}\ln\left(g(\theta)\right)+r\cdot G''(\theta+s)-r\cdot G''(\theta)$$

This second derivative is negative for all θ if $g''(\theta)\leq 0$ because $G''(\theta+s)-G''(\theta)<0$. The priors $g(\theta)$ that were considered herein were of two forms: $g(\theta)$ = θ^{-m} with $m > 0$ or $g(\theta)$ = $exp(-\theta/n)$ with $n > 0$.

$$\frac{d^2}{d\theta^2}\left(\theta^{-m}\right)=m(m-1)\theta^{m-2}$$

$$\frac{d^2}{d\theta^2}\left(e^{-\theta/n}\right)=\frac{1}{n^2}e^{-\theta/n}$$

The conditional posterior is unimodal if a uniform prior is used (*i.e.* $m = 0$), but may not be otherwise. The other priors could potentially have several critical points.

The second conditional posterior can rewritten as $f(\rho)\propto g(\rho)-r\cdot G(\rho)+\rho\cdot\left(-r\cdot\ln\lambda+\sum c_i\right)$, in which $g(\rho)$ is the prior on ρ conditional on λ.

$$\frac{d}{d\rho}f(\rho)=\frac{d}{d\rho}g(\rho)-r\cdot G'(\rho)-r\cdot\ln\lambda+\sum c_i$$

$$\frac{d^2}{d\rho^2}f(\rho)=\frac{d^2}{d\rho^2}g(\rho)-r\cdot G''(\rho)$$

This second derivative is negative for all ρ if $g''(\rho)\leq 0$ because $G''(\rho)>0$. In this thesis, $g(\rho)$ is the semi-infinite uniform prior, which results in a unimodal conditional posterior.

Appendix B
Derivation of Joint and Conditional Posterior Distribution Functions

This appendix contains the derivations of the joint and conditional posterior distribution functions that are obtained for each hierarchical Bayesian model developed in this thesis. These include (1) models for replicate recovery data, (2) models for paired matrix spike recovery data, (3) models for replicate concentration data, (4) models for single log-reduction estimates, and (5) models for temporally distributed concentration data.

B.1 Models for Replicate Recovery Data

This section presents derivations of the joint and conditional posterior distribution functions for three hierarchical models addressing random errors in replicate recovery measurements (Sections 4.2.2, 4.2.3, and 4.2.4). These are the beta-binomial recovery model (Appendix B.1.1), the beta-Poisson recovery model (Appendix B.1.2) and the negative binomial recovery model (Appendix B.1.3). The first two models can be combined (as shown in Figure B.1) if replicate recovery data have a mixture of precisely known and Poisson-distributed seed doses. The conditional posterior distribution functions for the beta-binomial recovery model (Appendix B.1.1) are used in Gibbs sampling algorithms in Sections 4.4.3, 4.4.4, 5.3.4, and 7.2. The posterior and conditional posteriors for the beta-Poisson and negative binomial recovery models are not used in this thesis.

B.1.1 Beta-binomial Recovery Model

Definitions

$x_i \sim BINOMIAL(n_i, p_i)$	number of particles observed in the i^{th} sample
$p_i \sim BETA(a,b)$	analytical recovery of the i^{th} sample

n_i the precisely known number of particles seeded into the i^{th} sample

a,b beta distribution shape parameters

r number of replicate seeded samples

Joint Distribution

$$f(\{x_i\},\{p_i\} \mid \{n_i\},a,b) = \prod_{i=1}^{r}\left[\frac{n_i!}{x_i!(n_i - x_i)!} p_i^{x_i}(1-p_i)^{n_i - x_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p_i^{a-1}(1-p_i)^{b-1}\right]$$

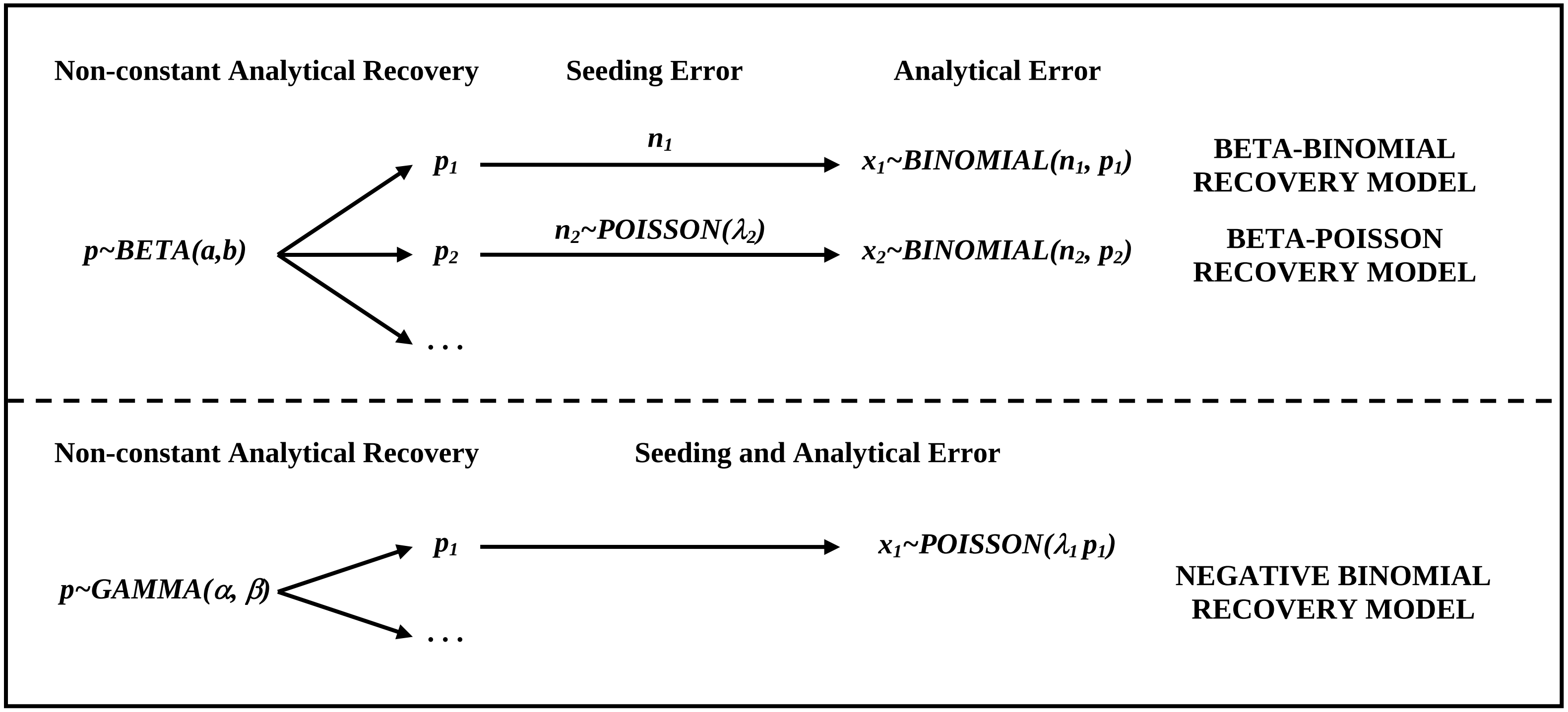

Non-constant Analytical Recovery
Seeding Error
Analytical Error
p~$BETA(a,b)$
p_1
n_1
x_1~$BINOMIAL(n_1, p_1)$
BETA-BINOMIAL RECOVERY MODEL
p_2
n_2~$POISSON(\lambda_2)$
x_2~$BINOMIAL(n_2, p_2)$
BETA-POISSON RECOVERY MODEL
. . .
Non-constant Analytical Recovery
Seeding and Analytical Error
p~$GAMMA(\alpha, \beta)$
p_1
x_1~$POISSON(\lambda_1 p_1)$
NEGATIVE BINOMIAL RECOVERY MODEL
. . .

Figure B.1: Probabilistic models for random errors in replicate recovery data

The three models for random errors in replicate recovery data are distinguished by the distribution chosen to represent non-constant analytical recovery and by the seeding methodology. Because the beta-binomial and beta-Poisson recovery models share beta-distributed analytical recovery, they can be combined into a single hierarchical model as shown. Recovery (*p*) is regarded as a probability in the models that use beta-distributed non-constant analytical recovery, which enables seeding and analytical errors to be considered separately. In these models, *n* is the number of seeded particles and *x* is the number of observed particles. λ is the expected number of seeded particles when the sample is seeded with a specified volume withdrawn from a stock of known concentration. In the negative binomial model, in which recovery is regarded as a rate that can exceed 100% due to counting errors, the seeding and analytical errors are combined into a single distribution.

Joint Posterior Distribution

$$f\left(a,b,\{p_i\}|\{x_i\},\{n_i\}\right)\propto g(a,b)\cdot\prod_{i=1}^{r}\left[\frac{n_i!}{x_i!(n_i-x_i)!}p_i^{x_i}(1-p_i)^{n_i-x_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p_i^{a-1}(1-p_i)^{b-1}\right]$$

Conditional Posterior Distributions

$$Dfc(a)\propto g(a|b)\cdot\left(\frac{\Gamma(a+b)}{\Gamma(a)}\right)^{r}\left(\prod_{i=1}^{r}p_i\right)^{a}$$

$$Dfc(b)\propto g(b|a)\cdot\left(\frac{\Gamma(a+b)}{\Gamma(b)}\right)^{r}\left(\prod_{i=1}^{r}(1-p_i)\right)^{b}$$

$$Dfc(p_i)\propto p_i^{x_i+a-1}(1-p_i)^{n_i-x_i+b-1}\rightarrow p_i\sim BETA(x_i+a,n_i-x_i+b)$$

Notes

Development of a Gibbs sampling algorithm is complicated by the conditional posterior distributions of a and b. An algorithm to sample from these distributions (approximately) is addressed in Section 4.4.3.

B.1.2 Beta-Poisson Recovery Model

Definitions

$n_i\sim POISSON(\lambda_i)$ number of particles seeded into the i^{th} sample

$x_i\sim BINOMIAL(n_i,p_i)$ number of particles observed in the i^{th} sample

$p_i\sim BETA(a,b)$ analytical recovery of the i^{th} sample

λ_i the expected number of particles seeded into the i^{th} sample ($\lambda_I = c_sV_{si}$)

c_s,V_{si} the particle stock concentration and the volume seeded into the i^{th} sample

a,b beta distribution shape parameters

r number of replicate seeded samples

Joint Distribution

$$f\left(\{n_i\},\{x_i\},\{p_i\}|\{\lambda_i\},a,b\right)=\prod_{i=1}^{r}\left[\frac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!}\right]\left[\frac{n_i!}{x_i!(n_i-x_i)!}p_i^{x_i}(1-p_i)^{n_i-x_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p_i^{a-1}(1-p_i)^{b-1}\right]$$

Joint PosteriorDistribution

$$f\left(a,b,\{n_i\},\{p_i\}|\{x_i\},\{\lambda_i\}\right)\propto$$

$$g(a,b)\cdot\prod_{i=1}^{r}\left[\frac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!}\right]\left[\frac{n_i!}{x_i!(n_i-x_i)!}p_i^{x_i}(1-p_i)^{n_i-x_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p_i^{a-1}(1-p_i)^{b-1}\right]$$

Conditional Posterior Distributions

$$Dfc(a) \propto g(a \mid b) \cdot \left(\frac{\Gamma(a+b)}{\Gamma(a)} \right)^r \left(\prod_{i=1}^{r} p_i \right)^a$$

$$Dfc(b) \propto g(b \mid a) \cdot \left(\frac{\Gamma(a+b)}{\Gamma(b)} \right)^r \left(\prod_{i=1}^{r} (1 - p_i) \right)^b$$

$$Dfc(n_i) \propto \frac{\lambda_i^{n_i}}{(n_i - x_i)!} (1 - p_i)^{n_i} \rightarrow n_i - x_i \sim POISSON(\lambda_i (1 - p_i))$$

$$Dfc(p_i) \propto p_i^{x_i + a - 1} (1 - p_i)^{n_i - x_i + b - 1} \rightarrow p_i \sim BETA(x_i + a, n_i - x_i + b)$$

Notes

Development of a Gibbs sampling algorithm is complicated by the conditional posterior distributions of *a* and *b*. An algorithm to sample from these distributions (approximately) is addressed in Section 4.4.3. This model can be used when replicate recovery data have a mixture of precisely known and Poisson-distributed seed doses simply by fixing the value of n_i for samples with known seed doses. It is possible to combine the Poisson and binomial distributions into x_i~Poisson($\lambda_i p_i$), but the resulting conditional posterior distribution for p_i is complicated.

B.1.3 Negative Binomial Recovery Model

Definitions

$x_i \sim POISSON(\lambda_i p_i)$ number of particles observed in the *i*th sample

$p_i \sim GAMMA(\alpha, \beta)$ analytical recovery of the *i*th sample

λ_i the expected number of particles seeded into the *i*th sample ($\lambda_i = c_s V_s$)

c_s, V_s the particle stock concentration and the volume seeded into the *i*th sample

α, β gamma distribution shape and scale parameters

r number of replicate seeded samples

Joint Distribution

$$f(\{x_i\}, \{p_i\} \mid \{\lambda_i\}, \alpha, \beta) = \prod_{i=1}^{r} \left[\frac{e^{-(\lambda_i p_i)} (\lambda_i p_i)^{x_i}}{x_i!} \right] \left[\frac{1}{\beta^{\alpha} \Gamma(\alpha)} p_i^{\alpha - 1} e^{-p_i / \beta} \right]$$

Joint Posterior Distribution

$$f(\alpha, \beta, \{p_i\} \mid \{x_i\}, \{\lambda_i\}) \propto g(\alpha, \beta) \cdot \prod_{i=1}^{r} \left[\frac{e^{-(\lambda_i p_i)} (\lambda_i p_i)^{x_i}}{x_i!} \right] \left[\frac{1}{\beta^{\alpha} \Gamma(\alpha)} p_i^{\alpha - 1} e^{-p_i / \beta} \right]$$

Conditional Posterior Distributions

$$Dfc(\alpha) \propto g(\alpha \mid \beta) \cdot (\Gamma(\alpha))^{-r} \left(\frac{1}{\beta} \prod_{i=1}^{r} p_i \right)^{\alpha}$$

$$Dfc(\beta) \propto g(\beta \mid \alpha) \cdot \left(\frac{1}{\beta} \right)^{r\alpha} e^{-\left(\sum p_i/\beta\right)}$$

$$Dfc(p_i) \propto p_i^{x_i+\alpha-1} e^{-p_i(\lambda_i+1/\beta)} \rightarrow \frac{p_i}{\beta}(\lambda_i \beta + 1) \sim GAMMA(x_i + \alpha, 1)$$

Notes

Development of a Gibbs sampling algorithm is complicated by the conditional posterior distribution of α. An acceptance-rejection algorithm to sample from a distribution of this form (approximately) is addressed in Section 6.3.1. If the conditional prior $g(\beta \mid \alpha)$ is either an improper semi-infinite uniform prior or a gamma distribution, then the conditional posterior for β^{-1} is gamma-distributed.

B.2 MODELS FOR PAIRED MATRIX SPIKE RECOVERY DATA

This section presents derivations of the joint and conditional posterior distribution functions for a model of paired matrix spike (oo)cyst recovery data (modified from Section 4.2.6) that are collected in accordance with Method 1623 (USEPA, 2005). Two models are presented: one in which the seed dose is precisely known (Appendix B.2.1) and one in which it is Poisson-distributed (Appendix B.2.2). The models presented in this section are simplified from the model presented in Figure 4.2 and use slightly different parameter definitions. The indigenous (oo)cyst concentrations in the two samples (c_1, c_2) are assumed to be equal because they are split or concurrent samples. Because matrix spike recoveries are assumed to be representative of the recovery of indigenous (oo)cysts, the analytical recovery of indigenous and seeded (oo)cysts in the matrix spike sample (p_2, p_2^*) are assumed to be equal, and the analytical error associated with the seeded and indigenous (oo)cysts is combined into a single binomial distribution. It is also assumed that the recovery in the unspiked field sample and the matrix spike sample (p_1, p_2) are equal, although some random variability in recovery among paired samples may exist. Matrix spike duplicates are not addressed, but they would have an equivalent model to the matrix spike sample (*e.g.* with subscript 3 on all parameters).

B.2.1 Matrix Spike Recovery Model with Precise Seed Dose

Definitions

$n_1 \sim POISSON(cV_1)$ number of (oo)cysts in unspiked field sample

$n_2 \sim POISSON(cV_2) + T$ number of (oo)cysts in matrix spike sample

$x_1 \sim BINOMIAL(n_1, p)$ observed count of (oo)cysts in unspiked field sample

$x_2 \sim BINOMIAL(n_2, p)$ observed count of (oo)cysts in matrix spike sample

c concentration of indigenous (oo)cysts in the source water

V_1,V_2 sample volumes of the unspiked field sample and matrix spike sample respectively

T the precisely known number of (oo)cysts seeded into the matrix spike sample

p (oo)cyst recovery in the unspiked field sample and the matrix spike sample

Joint Distribution

$$f(n_1, n_2, x_1, x_2 \mid c, V_1, V_2, T, p) =$$

$$\left[\frac{e^{-cV_1}(cV_1)^{n_1}}{n_1!}\right]\left[\frac{e^{-cV_2}(cV_2)^{n_2-T}}{(n_2-T)!}\right]\left[\frac{n_1!}{x_1!(n_1-x_1)!}p^{x_1}(1-p)^{n_1-x_1}\right]\left[\frac{n_2!}{x_2!(n_2-x_2)!}p^{x_2}(1-p)^{n_2-x_2}\right]$$

Joint Posterior Distribution

$$f(c, n_1, n_2, p \mid x_1, x_2, V_1, V_2, T) \propto$$

$$g(c,p)\cdot\left[\frac{e^{-cV_1}(cV_1)^{n_1}}{n_1!}\right]\left[\frac{e^{-cV_2}(cV_2)^{n_2-T}}{(n_2-T)!}\right]\left[\frac{n_1!}{x_1!(n_1-x_1)!}p^{x_1}(1-p)^{n_1-x_1}\right]\left[\frac{n_2!}{x_2!(n_2-x_2)!}p^{x_2}(1-p)^{n_2-x_2}\right]$$

Conditional Posterior Distributions

* Assume $g(c,p) = g(c) \cdot g(p)$ (*i.e.* that c and p are independent *a priori*).

$$Dfc(c) \propto g(c) \cdot c^{n_1+n_2-T} e^{-c(V_1+V_2)}$$

$$Dfc(n_1) \propto \frac{(cV_1)^{n_1}}{(n_1-x_1)!}(1-p)^{n_1} \rightarrow n_1 - x_1 \sim POISSON(cV_1(1-p))$$

$$Dfc(n_2) \propto \frac{n_2!}{(n_2-T)!(n_2-x_2)!}(cV_2(1-p))^{n_2}$$

$$Dfc(p) \propto g(p) \cdot p^{x_1+x_2}(1-p)^{n_1+n_2-x_1-x_2}$$

Notes

There is an alternative parameterization in which n_2~Poisson(cV_2) and x_2~Binomial(n_2+T,p). The parameter n_2 becomes the number of indigenous (oo)cysts rather than the total number of (oo)cysts in the matrix spike sample. The results are the same except that every occurrence of n_2 above is replaced with n_2+T. Development of a Gibbs sampling algorithm is complicated by the conditional posterior distribution of n_2 and is not addressed herein. If the prior for concentration is either an improper uniform prior or a gamma distribution, then its conditional posterior is gamma-distributed. If a beta-distributed prior (including the standard uniform distribution) is used for the probability of recovery, then its conditional posterior is beta-distributed.

B.2.2 Matrix Spike Recovery Model with Poisson-distributed Seed Doses

Definitions

$n_1 \sim POISSON(cV_1)$ number of (oo)cysts in unspiked field sample
$n_2 \sim POISSON(cV_2 + T)$ number of (oo)cysts in matrix spike sample
$x_1 \sim BINOMIAL(n_1, p)$ observed count of (oo)cysts in unspiked field sample
$x_2 \sim BINOMIAL(n_2, p)$ observed count of (oo)cysts in matrix spike sample
c concentration of indigenous (oo)cysts in the source water
V_1, V_2 sample volumes of the unspiked field sample and matrix spike sample respectively
T the expected number of (oo)cysts seeded into the matrix spike sample ($T = c^*V^*$)
c^*, V^* the (oo)cyst stock concentration and the volume seeded into the matrix spike sample
p (oo)cyst recovery in the unspiked field sample and the matrix spike sample

Joint Distribution

$$f(n_1, n_2, x_1, x_2 \mid c, V_1, V_2, T, p) =$$
$$\left[\frac{e^{-cV_1}(cV_1)^{n_1}}{n_1!}\right]\left[\frac{e^{-(cV_2+T)}(cV_2+T)^{n_2}}{n_2!}\right]\left[\frac{n_1!}{x_1!(n_1-x_1)!}p^{x_1}(1-p)^{n_1-x_1}\right]\left[\frac{n_2!}{x_2!(n_2-x_2)!}p^{x_2}(1-p)^{n_2-x_2}\right]$$

Joint Posterior Distribution

$$f(c, n_1, n_2, p \mid x_1, x_2, V_1, V_2, T) \propto g(c, p) \cdot$$
$$\left[\frac{e^{-cV_1}(cV_1)^{n_1}}{n_1!}\right]\left[\frac{e^{-(cV_2+T)}(cV_2+T)^{n_2}}{n_2!}\right]\left[\frac{n_1!}{x_1!(n_1-x_1)!}p^{x_1}(1-p)^{n_1-x_1}\right]\left[\frac{n_2!}{x_2!(n_2-x_2)!}p^{x_2}(1-p)^{n_2-x_2}\right]$$

Conditional Posterior Distributions

* Assume $g(c,p) = g(c) \cdot g(p)$ (*i.e.* that c and p are independent *a priori*).

$$Dfc(c) \propto g(c) \cdot c^{n_1}(cV_2 + T)^{n_2} e^{-c(V_1+V_2)}$$

$$Dfc(n_1) \propto \frac{(cV_1)^{n_1}}{(n_1 - x_1)!}(1-p)^{n_1} \rightarrow n_1 - x_1 \sim POISSON(cV_1(1-p))$$

$$Dfc(n_2) \propto \frac{(cV_2 + T)^{n_2}}{(n_2 - x_2)!}(1-p)^{n_2} \rightarrow n_2 - x_2 \sim POISSON((cV_2 + T)(1-p))$$

$$Dfc(p) \propto g(p) \cdot p^{x_1+x_2}(1-p)^{n_1+n_2-x_1-x_2}$$

Notes

Development of a Gibbs sampling algorithm is complicated by the conditional posterior distribution of c and is not addressed herein. If a beta-distributed prior (including the standard uniform distribution) is used for the probability of recovery, then its conditional posterior is beta-distributed.

B.3 MODELS FOR REPLICATE CONCENTRATION DATA

This section presents derivations of the joint and conditional posterior distribution functions for three hierarchical models addressing random errors in replicate enumeration data (Section 5.1). These are the beta-Poisson enumeration model (Appendix B.3.1), the internal seed enumeration model (Appendix B.3.2) and the negative binomial enumeration model (Appendix B.3.3). The first two models can be combined (as shown in Figure B.2). The conditional posterior distribution functions derived in this section for the beta-Poisson and negative binomial enumeration models are used in Gibbs sampling algorithms in Sections 5.3.2 and 5.3.3 (except that the same non-constant analytical recovery distribution parameters are used for all replicate samples in those sections).

B.3.1 Beta-Poisson Enumeration Model

Definitions

$n_i \sim POISSON(cV_i)$	number of particles in the i^{th} replicate
$x_i \sim BINOMIAL(n_i, p_i)$	number of particles observed in the i^{th} replicate
$p_i \sim BETA(a_i, b_i)$	analytical recovery of the i^{th} replicate
c	the particle concentration in the source
V_i	the sample volume of the i^{th} replicate
a_i,b_i	beta distribution shape parameters for the i^{th} replicate (often not sample-specific)
r	number of replicate samples

Joint Distribution

$$f(\{n_i\},\{x_i\},\{p_i\} \mid c,\{V_i\},\{a_i\},\{b_i\}) =$$

$$\prod_{i=1}^{r}\left[\frac{e^{-cV_i}(cV_i)^{n_i}}{n_i!}\right]\left[\frac{n_i!}{x_i!(n_i-x_i)!}p_i^{x_i}(1-p_i)^{n_i-x_i}\right]\left[\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}p_i^{a_i-1}(1-p_i)^{b_i-1}\right]$$

Joint Posterior Distribution

Assume uniform prior distributions

$$f(c,\{n_i\},\{p_i\} \mid \{x_i\},\{V_i\},\{a_i\},\{b_i\}) \propto$$

$$g(c)\cdot\prod_{i=1}^{r}\left[\frac{e^{-cV_i}(cV_i)^{n_i}}{n_i!}\right]\left[\frac{n_i!}{x_i!(n_i-x_i)!}p_i^{x_i}(1-p_i)^{n_i-x_i}\right]\left[\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}p_i^{a_i-1}(1-p_i)^{b_i-1}\right]$$

Conditional Posterior Distributions

$$Dfc(c) \propto g(c)\cdot c^{\sum n_i}e^{-c\sum V_i}$$

$$Dfc(n_i) \propto \frac{(cV_i)^{n_i}}{(n_i-x_i)!}(1-p_i)^{n_i} \rightarrow n_i - x_i \sim POISSON(cV_i(1-p_i))$$

$$Dfc(p_i) \propto p_i^{x_i+a_i-1}(1-p_i)^{n_i-x_i+b_i-1} \rightarrow p_i \sim BETA(x_i+a_i, n_i-x_i+b_i)$$

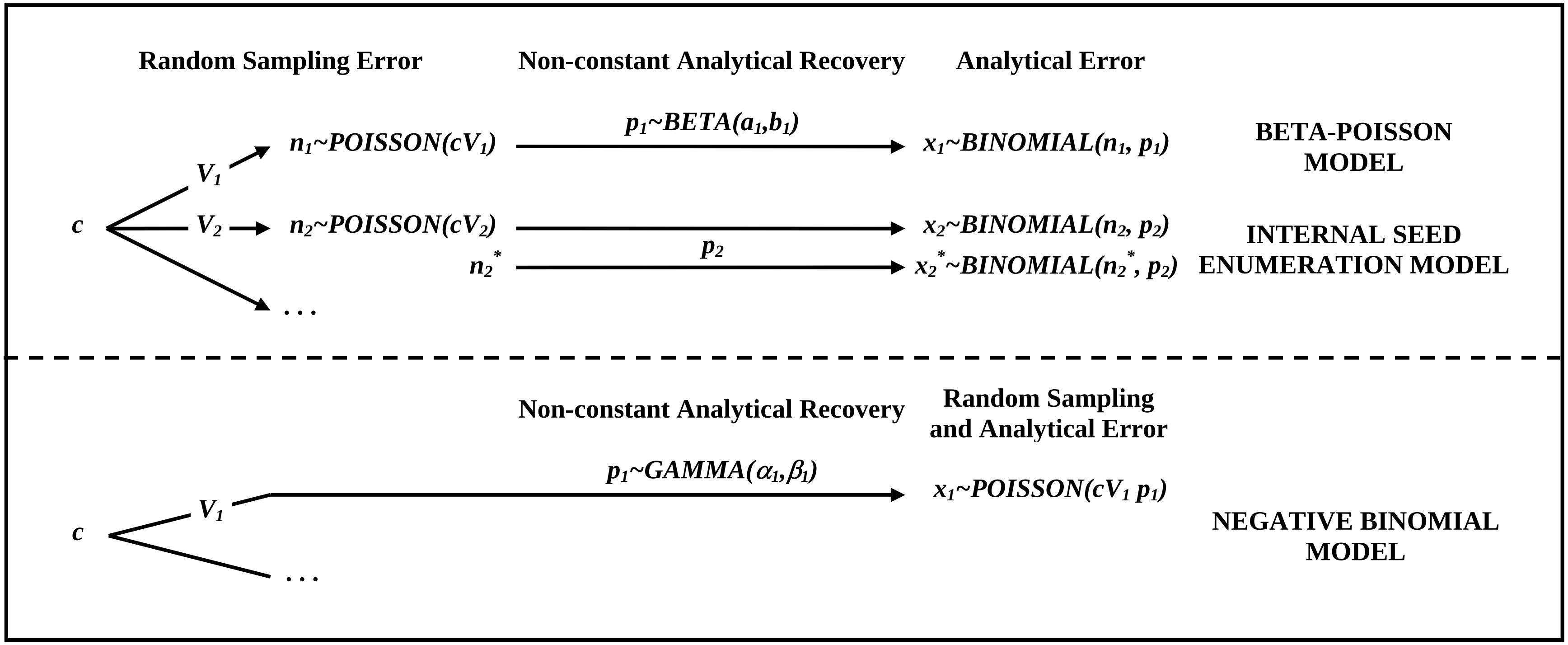

Random Sampling Error
Non-constant Analytical Recovery
Analytical Error
c
V_1
V_2
n_1~POISSON(cV_1)
p_1~BETA(a_1,b_1)
x_1~BINOMIAL(n_1, p_1)
BETA-POISSON MODEL
n_2~POISSON(cV_2)
x_2~BINOMIAL(n_2, p_2)
p_2
n_2^*
x_2^*~BINOMIAL(n_2^*, p_2)
INTERNAL SEED ENUMERATION MODEL
. . .
Non-constant Analytical Recovery
Random Sampling and Analytical Error
p_1~GAMMA(α_1,β_1)
c
V_1
x_1~POISSON(cV_1 p_1)
NEGATIVE BINOMIAL MODEL
. . .

Figure B.2: Probabilistic models for random errors in replicate enumeration data

The three models for random errors in replicate enumeration data are distinguished by the available information about analytical recovery. Recovery (*p*) is regarded as a probability in both the beta-Poisson and internal seed recovery models, which enables random sampling and analytical errors to be considered separately and allows the two models to be combined into a single hierarchical model as shown. In these models, *n* is the number of particles in the sample and *x* is the number of observed particles. n^* is the precisely known number of seeded particles, of which x^* are observed, in the internal seed enumeration model. In the negative binomial model, in which recovery is regarded as a rate that can exceed 100% due to counting errors, the random sampling and analytical errors are combined into a single distribution.

Notes

This model is generalized so that any sample from the same source is regarded as a replicate: the volume and recovery distribution parameters are potentially sample-specific. The special case where all samples have the same recovery parameters is presented in Section 5.3.2. It is possible to combine the Poisson and binomial distributions into x_i~Poisson($cV_{ii}p_i$), but the resulting conditional posterior distribution for p_i is complicated. If the prior for concentration is either an improper semi-infinite uniform prior or a gamma distribution, then its conditional posterior is gamma-distributed.

B.3.2 Internal Seed Enumeration Model

Definitions

$n_i \sim POISSON(cV_i)$ number of indigenous particles in the i^{th} replicate

$x_i \sim BINOMIAL(n_i, p_i)$ number of indigenous particles observed in the i^{th} replicate

$x_i^* \sim BINOMIAL(n_i^*, p_i)$ number of seeded particles observed in the i^{th} replicate

c the particle concentration in the source

V_i the sample volume of the i^{th} replicate

n_i^* the precisely known number of seeded particles in the i^{th} replicate

x_i^* the number of seeded particles observed in the i^{th} replicate

r number of replicate samples

Joint Distribution

$$f\left(\{n_i\},\{x_i\},\{x_i^*\} \mid c,\{V_i\},\{p_i\},\{n_i^*\}\right)=$$

$$\prod_{i=1}^{r}\left[\frac{e^{-cV_i}(cV_i)^{n_i}}{n_i!}\right]\left[\frac{n_i!}{x_i!(n_i-x_i)!}p_i^{x_i}(1-p_i)^{n_i-x_i}\right]\left[\frac{n_i^*!}{x_i^*!(n_i^*-x_i^*)!}p_i^{x_i^*}(1-p_i)^{n_i^*-x_i^*}\right]$$

Joint Posterior Distribution

$$f\left(c,\{n_i\},\{p_i\} \mid \{x_i\},\{x_i^*\},\{V_i\},\{n_i^*\}\right)\propto$$

$$g(c,\{p_i\})\cdot\prod_{i=1}^{r}\left[\frac{e^{-cV_i}(cV_i)^{n_i}}{n_i!}\right]\left[\frac{n_i!}{x_i!(n_i-x_i)!}p_i^{x_i}(1-p_i)^{n_i-x_i}\right]\left[\frac{n_i^*!}{x_i^*!(n_i^*-x_i^*)!}p_i^{x_i^*}(1-p_i)^{n_i^*-x_i^*}\right]$$

Conditional Posterior Distributions

$$Dfc(c)\propto g(c)\cdot c^{\sum n_i}e^{-c\sum V_i}$$

$$Dfc(n_i)\propto\frac{(cV_i)^{n_i}}{(n_i-x_i)!}(1-p_i)^{n_i}\rightarrow n_i-x_i\sim POISSON(cV_i(1-p_i))$$

$$Dfc(p_i)\propto g(p_i)\cdot p_i^{x_i+x_i^*}(1-p_i)^{n_i+n_i^*-x_i-x_i^*}$$

Notes

This model is generalized so that any sample from the same source is regarded as a replicate: the volume and number of seeded particles are potentially sample-specific. If the prior for concentration is either an improper semi-infinite uniform prior or a gamma distribution, then its conditional posterior is gamma-distributed. If a beta-distributed prior (including the standard uniform distribution) is used for the probability of recovery, then its conditional posterior is beta-distributed (as discussed in Section 5.3.4). If a potentially sample-specific beta-distributed prior (p_i~Beta(a_i^*,b_i^*)) is used, then the conditional posterior for the probability of recovery is beta-distributed with the parameters ($x_i + x_i^* + a_i^*$, $n_i - x_i + n_i^* - x_i^* + b_i^*$). The standard uniform prior is the special case in which $a_i^* = 1$, $b_i^* = 1$. Accordingly, it is equivalent to incorporate internal seed recovery data and beta-distributed priors for the probability of recovery into enumeration data analysis as presented in this section or as a special case of the beta-Poisson model with the sample-specific beta distribution parameters $a_i = x_i^* + a_i^*$, $b_i = n_i^* - x_i^* + b_i^*$. In this case, the beta distribution becomes an informative prior upon the probability of recovery for indigenous particles rather than a description of non-constant analytical recovery.

B.3.3 Gamma-distributed Recovery

Definitions

$x_i \sim POISSON(cV_i p_i)$ number of particles observed in the i^{th} replicate
$p_i \sim GAMMA(\alpha_i, \beta_i)$ analytical recovery of the i^{th} replicate
c the particle concentration in the source
V_i the sample volume of the i^{th} replicate
α_i,β_i gamma distribution shape and scale parameters for the i^{th} replicate
r number of replicate samples

Joint Distribution

$$f(\{x_i\},\{p_i\} \mid c,\{V_i\},\{\alpha_i\},\{\beta_i\}) = \prod_{i=1}^{r}\left[\frac{e^{-cV_i p_i}(cV_i p_i)^{x_i}}{x_i!}\right]\left[\frac{1}{\beta_i^{\alpha_i}\Gamma(\alpha_i)}p_i^{\alpha_i-1}e^{-p_i/\beta_i}\right]$$

Joint Posterior Distribution

$$f(c,\{p_i\} \mid \{x_i\},\{V_i\},\{\alpha_i\},\{\beta_i\}) = g(c)\cdot\prod_{i=1}^{r}\left[\frac{e^{-cV_i p_i}(cV_i p_i)^{x_i}}{x_i!}\right]\left[\frac{1}{\beta_i^{\alpha_i}\Gamma(\alpha_i)}p_i^{\alpha_i-1}e^{-p_i/\beta_i}\right]$$

Conditional Posterior Distributions

$$Dfc(c) \propto g(c)\cdot c^{\sum x_i} e^{-c\sum V_i p_i}$$

$$Dfc(p_i) \propto p_i^{x_i+\alpha_i-1}e^{-p_i(cV_i+1/\beta_i)} \rightarrow \frac{p_i}{\beta_i}(cV_i\beta_i+1) \sim GAMMA(x_i+\alpha_i,1)$$

Notes

This model is generalized so that any sample from the same source is regarded as a replicate: the volume and recovery distribution parameters are potentially sample-specific. The special case where all samples have the same recovery parameters is presented in Section 5.3.2. It is possible to combine the Poisson and gamma distributions into $x_i \sim \text{NegativeBinomial}(\alpha_i,(cV_i\beta_i)^{-1})$, but the resulting conditional posterior distribution for c is complicated (as discussed in Section 5.3.2). If the prior for concentration is either an improper semi-infinite uniform prior or a gamma distribution, then its conditional posterior is gamma-distributed.

B.4 MODELS FOR SINGLE LOG-REDUCTION ESTIMATES

This section presents derivations of the joint and conditional posterior distribution functions for two hierarchical log-reduction models. These models are based upon the models presented in Appendix B.3: two sets of concentration data are related by log-reduction (ρ) using the relationship $c_2 = c_1 \cdot 10^{-\rho}$ (as discussed in Section 5.4.3. Use of Bayes' theorem with these models describes uncertainty in a single log-reduction estimate, not variability in log-reduction over time (or among multiple subsets of data). The beta-Poisson and negative binomial log-reduction models are presented in Appendices B.4.1 and A.4.2 respectively. Gibbs sampling using the beta-Poisson log-reduction model is discussed in Section 5.4.3.

B.4.1 Beta-Poisson log-Reduction Model

Definitions

$n_{1k} \sim POISSON(cV_{1k})$ number of particles in the kth replicate of the initial conc. data

$n_{2k} \sim POISSON(c \cdot 10^{-\rho} \cdot V_{2k})$ number of particles in the kth replicate of the final conc. data

$x_{1k} \sim BINOMIAL(n_{1k}, p_{1k})$ number of particles observed in the kth replicate of the initial concentration data

$x_{2k} \sim BINOMIAL(n_{2k}, p_{2k})$ number of particles observed in the kth replicate of the final concentration data

$p_{1k} \sim BETA(a_{1k}, b_{1k})$ analytical recovery of the kth replicate of the initial conc. data

$p_{2k} \sim BETA(a_{2k}, b_{2k})$ analytical recovery of the kth replicate of the final conc. data

c initial particle concentration

ρ log-reduction

V_{1k} the sample volume of the kth replicate in the initial concentration data

V_{2k} the sample volume of the kth replicate in the final concentration data

a_{1k},b_{1k} beta distribution shape parameters for the kth replicate in the initial concentration data

a_{2k},b_{2k} beta distribution shape parameters for the kth replicate in the final concentration data

r_1 number of replicate samples in the initial concentration dataset

r_2 number of replicate samples in the final concentration dataset

Joint Distribution

$$f\left(\{n_{1k}\},\{n_{2k}\},\{x_{1k}\},\{x_{2k}\},\{p_{1k}\},\{p_{2k}\}\,|\,c,\rho,\{V_{1k}\},\{V_{2k}\},\{a_{1k}\},\{a_{2k}\},\{b_{1k}\},\{b_{2k}\}\right)=$$

$$\left[\prod_{k=1}^{r_1}\left[\frac{e^{-cV_{1k}}\left(cV_{1k}\right)^{n_{1k}}}{n_{1k}!}\right]\left[\frac{n_{1k}!}{x_{1k}!\left(n_{1k}-x_{1k}\right)!}\,p_{1k}^{x_{1k}}\left(1-p_{1k}\right)^{n_{1k}-x_{1k}}\right]\left[\frac{\Gamma\left(a_{1k}+b_{1k}\right)}{\Gamma\left(a_{1k}\right)\Gamma\left(b_{1k}\right)}\,p_{1k}^{a_{1k}-1}\left(1-p_{1k}\right)^{b_{1k}-1}\right]\right]\cdot$$

$$\left[\prod_{k=1}^{r_2}\left[\frac{e^{-10^{-\rho}cV_{2k}}\left(10^{-\rho}cV_{2k}\right)^{n_{2k}}}{n_{2k}!}\right]\left[\frac{n_{2k}!\cdot p_{2k}^{x_{2k}}\left(1-p_{2k}\right)^{n_{2k}-x_{2k}}}{x_{2k}!\left(n_{2k}-x_{2k}\right)!}\right]\left[\frac{\Gamma\left(a_{2k}+b_{2k}\right)}{\Gamma\left(a_{2k}\right)\Gamma\left(b_{2k}\right)}\,p_{2k}^{a_{2k}-1}\left(1-p_{2k}\right)^{b_{2k}-1}\right]\right]$$

Joint Posterior Distribution

$$f\left(c,\rho,\{n_{1k}\},\{n_{2k}\},\{p_{1k}\},\{p_{2k}\}\,|\,\{x_{1k}\},\{x_{2k}\},\{V_{1k}\},\{V_{2k}\},\{a_{1k}\},\{a_{2k}\},\{b_{1k}\},\{b_{2k}\}\right)\propto g\left(c,\rho\right)\cdot$$

$$\left[\prod_{k=1}^{r_1}\left[\frac{e^{-cV_{1k}}\left(cV_{1k}\right)^{n_{1k}}}{n_{1k}!}\right]\left[\frac{n_{1k}!}{x_{1k}!\left(n_{1k}-x_{1k}\right)!}\,p_{1k}^{x_{1k}}\left(1-p_{1k}\right)^{n_{1k}-x_{1k}}\right]\left[\frac{\Gamma\left(a_{1k}+b_{1k}\right)}{\Gamma\left(a_{1k}\right)\Gamma\left(b_{1k}\right)}\,p_{1k}^{a_{1k}-1}\left(1-p_{1k}\right)^{b_{1k}-1}\right]\right]\cdot$$

$$\left[\prod_{k=1}^{r_2}\left[\frac{e^{-10^{-\rho}cV_{2k}}\left(10^{-\rho}cV_{2k}\right)^{n_{2k}}}{n_{2k}!}\right]\left[\frac{n_{2k}!\cdot p_{2k}^{x_{2k}}\left(1-p_{2k}\right)^{n_{2k}-x_{2k}}}{x_{2k}!\left(n_{2k}-x_{2k}\right)!}\right]\left[\frac{\Gamma\left(a_{2k}+b_{2k}\right)}{\Gamma\left(a_{2k}\right)\Gamma\left(b_{2k}\right)}\,p_{2k}^{a_{2k}-1}\left(1-p_{2k}\right)^{b_{2k}-1}\right]\right]$$

Conditional Posterior Distributions

$$Dfc(c)\propto g(c\,|\,\rho)\cdot\exp\left(-c\left(\sum_{k=1}^{r_1}V_{1k}+10^{-\rho}\cdot\sum_{k=1}^{r_2}V_{2k}\right)\right)\cdot c^\wedge\left(\sum_{k=1}^{r_1}n_{1k}+\sum_{k=1}^{r_2}n_{2k}\right)$$

$$Dfc(\rho)\propto g(\rho\,|\,c)\cdot\exp\left(-c\cdot 10^{-\rho}\cdot\sum_{k=1}^{r_2}V_{2k}\right)\cdot\left(10^{-\rho}\right)^\wedge\left(\sum_{k=1}^{r_2}n_{2k}\right)$$

$$Dfc\left(n_{1k}\right)\propto\frac{\left(cV_{1k}\right)^{n_{1k}}}{\left(n_{1k}-x_{1k}\right)!}\left(1-p_{1k}\right)^{n_{1k}}\rightarrow n_{1k}-x_{1k}\sim POISSON\left(cV_{1k}\left(1-p_{1k}\right)\right)$$

$$Dfc\left(n_{2k}\right)\propto\frac{\left(10^{-\rho}cV_{2k}\right)^{n_{2k}}}{\left(n_{2k}-x_{2k}\right)!}\left(1-p_{2k}\right)^{n_{2k}}\rightarrow n_{2k}-x_{2k}\sim POISSON\left(10^{-\rho}cV_{2k}\left(1-p_{2k}\right)\right)$$

$$Dfc\left(p_{1k}\right)\propto p_{1k}^{x_{1k}+a_{1k}-1}\left(1-p_{1k}\right)^{n_{1k}-x_{1k}+b_{1k}-1}\rightarrow p_{1k}\sim BETA\left(x_{1k}+a_{1k},n_{1k}-x_{1k}+b_{1k}\right)$$

$$Dfc\left(p_{2k}\right)\propto p_{2k}^{x_{2k}+a_{2k}-1}\left(1-p_{2k}\right)^{n_{2k}-x_{2k}+b_{2k}-1}\rightarrow p_{2k}\sim BETA\left(x_{2k}+a_{2k},n_{2k}-x_{2k}+b_{2k}\right)$$

Notes

This model is generalized so that any sample from the same source is regarded as a replicate: the volume and recovery distribution parameters are potentially sample-specific. The special case where all samples have the same recovery parameters is presented in Section 5.4.3. Use of internal seed recovery data is not discussed herein, but can be incorporated into this model (if a beta-distributed prior is used for each sample's probability of recovery) as described in Appendix B.3.2. If $g(c|\rho)$ is the improper semi-infinite uniform prior, then $c(\Sigma V_{1k} + 10^{-\rho}\cdot\Sigma V_{2k}) \sim GAMMA(\Sigma n_{1k} + \Sigma n_{2k} + 1,1)$ can be used to generate c. If $g(\rho\,|c)$ is the improper infinite uniform prior (because $-\infty < \rho < \infty$), then $10^{-\rho}\cdot c\cdot\Sigma V_{2k} \sim GAMMA(\Sigma n_{2k} + 1,1)$ can be used to generate ρ.

B.4.2 Negative Binomial log-Reduction Model

Definitions

$x_{1k} \sim POISSON(cV_{1k}\,p_{1k})$ number of particles observed in the k^{th} replicate of the initial concentration data

$x_{2k} \sim POISSON\left(c \cdot 10^{-\rho} \cdot V_{2k}\,p_{2k}\right)$ number of particles observed in the k^{th} replicate of the final concentration data

$p_{1k} \sim GAMMA(\alpha_{1k}, \beta_{1k})$ analytical recovery of the k^{th} replicate of the initial conc. data

$p_{2k} \sim GAMMA(\alpha_{2k}, \alpha_{2k})$ analytical recovery of the k^{th} replicate of the final conc. data

c initial particle concentration

ρ log-reduction

V_{1k} the sample volume of the k^{th} replicate in the initial concentration data

V_{2k} the sample volume of the k^{th} replicate in the final concentration data

α_{1k},β_{1k} gamma distribution parameters for the k^{th} replicate in the initial concentration data

α_{2k},β_{2k} gamma distribution parameters for the k^{th} replicate in the final concentration data

r_1 number of replicate samples in the initial concentration dataset

r_2 number of replicate samples in the final concentration dataset

Joint Distribution

$$f\left(\{x_{1k}\},\{x_{2k}\},\{p_{1k}\},\{p_{2k}\} \mid c,\rho,\{V_{1k}\},\{V_{2k}\},\{\alpha_{1k}\},\{\alpha_{2k}\},\{\beta_{1k}\},\{\beta_{2k}\}\right)=$$

$$\left[\prod_{k=1}^{r_1}\left[\frac{e^{-cV_{1k}p_{1k}}\left(cV_{1k}\,p_{1k}\right)^{x_{1k}}}{x_{1k}!}\right]\left[\frac{1}{\beta_{1k}^{\alpha_{1k}}\Gamma(\alpha_{1k})\Gamma(b_{1k})}\,p_{1k}^{\alpha_{1k}-1}e^{-p_{1k}/\beta_{1k}}\right]\right]\cdot$$

$$\left[\prod_{k=1}^{r_2}\left[\frac{e^{-10^{-\rho}cV_{2k}p_{2k}}\left(10^{-\rho}cV_{2k}\,p_{2k}\right)^{x_{2k}}}{x_{2k}!}\right]\left[\frac{1}{\beta_{2k}^{\alpha_{2k}}\Gamma(\alpha_{2k})}\,p_{2k}^{\alpha_{2k}-1}e^{-p_{2k}/\beta_{2k}}\right]\right]$$

Joint Posterior Distribution

$$f\left(c,\rho,\{p_{1k}\},\{p_{2k}\} \mid \{x_{1k}\},\{x_{2k}\},\{V_{1k}\},\{V_{2k}\},\{\alpha_{1k}\},\{\alpha_{2k}\},\{\beta_{1k}\},\{\beta_{2k}\}\right) \propto g(c,\rho)\cdot$$

$$\left[\prod_{k=1}^{r_1}\left[\frac{e^{-cV_{1k}p_{1k}}\left(cV_{1k}\,p_{1k}\right)^{x_{1k}}}{x_{1k}!}\right]\left[\frac{1}{\beta_{1k}^{\alpha_{1k}}\Gamma(\alpha_{1k})\Gamma(b_{1k})}\,p_{1k}^{\alpha_{1k}-1}e^{-p_{1k}/\beta_{1k}}\right]\right]\cdot$$

$$\left[\prod_{k=1}^{r_2}\left[\frac{e^{-10^{-\rho}cV_{2k}p_{2k}}\left(10^{-\rho}cV_{2k}\,p_{2k}\right)^{x_{2k}}}{x_{2k}!}\right]\left[\frac{1}{\beta_{2k}^{\alpha_{2k}}\Gamma(\alpha_{2k})}\,p_{2k}^{\alpha_{2k}-1}e^{-p_{2k}/\beta_{2k}}\right]\right]$$

Conditional Posterior Distributions

$$Dfc(c) \propto g(c \mid \rho)\cdot \exp\left(-c\left(\sum_{k=1}^{r_1} V_{1k} + 10^{-\rho}\cdot \sum_{k=1}^{r_2} V_{2k}\right)\right)\cdot c^\wedge\left(\sum_{k=1}^{r_1} x_{1k} + \sum_{k=1}^{r_2} x_{2k}\right)$$

$$Dfc(\rho) \propto g(\rho \mid c)\cdot \exp\left(-c\cdot 10^{-\rho}\cdot \sum_{k=1}^{r_2} V_{2k}\right)\cdot \left(10^{-\rho}\right)^\wedge\left(\sum_{k=1}^{r_2} x_{2k}\right)$$

$$Dfc(p_{1k}) \propto \exp\left(-p_{1k}\left(cV_{1k} + 1/\beta_{1k}\right)\right)\cdot p_{1k}^{x_{1k}+\alpha_{1k}-1} \rightarrow p_{1k}\left(cV_{1k} + 1/\beta_{1k}\right) \sim GAMMA\left(x_{1k}+\alpha_{1k},1\right)$$

$$Dfc(p_{2k}) \propto \exp\left(-p_{2k}\left(c\cdot 10^{-\rho}\cdot V_{2k} + 1/\beta_{2k}\right)\right)\cdot p_{2k}^{x_{2k}+\alpha_{2k}-1}$$

$$\rightarrow p_{2k}\left(c\cdot 10^{-\rho}\cdot V_{2k} + 1/\beta_{2k}\right) \sim GAMMA\left(x_{2k}+\alpha_{2k},1\right)$$

Notes

This model is generalized so that any sample from the same source is regarded as a replicate: the volume and recovery distribution parameters are potentially sample-specific. If $g(c|\rho)$ is the improper semi-infinite uniform prior, then $c(\Sigma V_{1k} + 10^{-\rho} \cdot \Sigma V_{2k}) \sim GAMMA(\Sigma n_{1k} + \Sigma n_{2k} + 1,1)$ can be used to generate c. If $g(\rho\,|c)$ is the improper infinite uniform prior (because $-\infty < \rho < \infty$), then $10^{-\rho} \cdot c \cdot \Sigma V_{2k} \sim GAMMA(\Sigma n_{2k} + 1,1)$ can be used to generate ρ.

B.5 Models for Temporally Distributed Concentration Data

This section presents derivations of the joint and conditional posterior distribution functions for three hierarchical models addressing random errors in temporally distributed enumeration data (Section 6.1.1). These are the beta-Poisson temporal concentration variability model (Appendix B.4.1), the internal seed temporal concentration variability model (Appendix B.4.2) and the negative binomial temporal concentration variability model (Appendix B.4.3). The first two models can be combined (as shown in Figure B.3). The conditional posterior distribution functions derived in this section are used in Gibbs sampling algorithms in Section 6.3.

B.5.1 Beta-Poisson Temporal Concentration Variability Model

Definitions

$c_i \sim GAMMA(\rho,\lambda)$ particle concentration during the ith sampling event

$n_{ik} \sim POISSON(c_i V_{ik})$ number of particles in the kth replicate of the ith sampling event

$x_{ik} \sim BINOMIAL(n_{ik}, p_{ik})$ number of particles observed in the kth replicate of the ith sampling event

$p_{ik} \sim BETA(a_{ik}, b_{ik})$ analytical recovery of the kth replicate of the ith sampling event

ρ,λ shape and scale parameters of the temporal concentration variability distribution

V_{ik} the sample volume of the kth replicate of the ith sampling event

a_{ik},b_{ik} beta distribution shape parameters for the kth replicate of the ith sampling event

r number of sampling events

r_i number of replicate samples in the ith sampling event

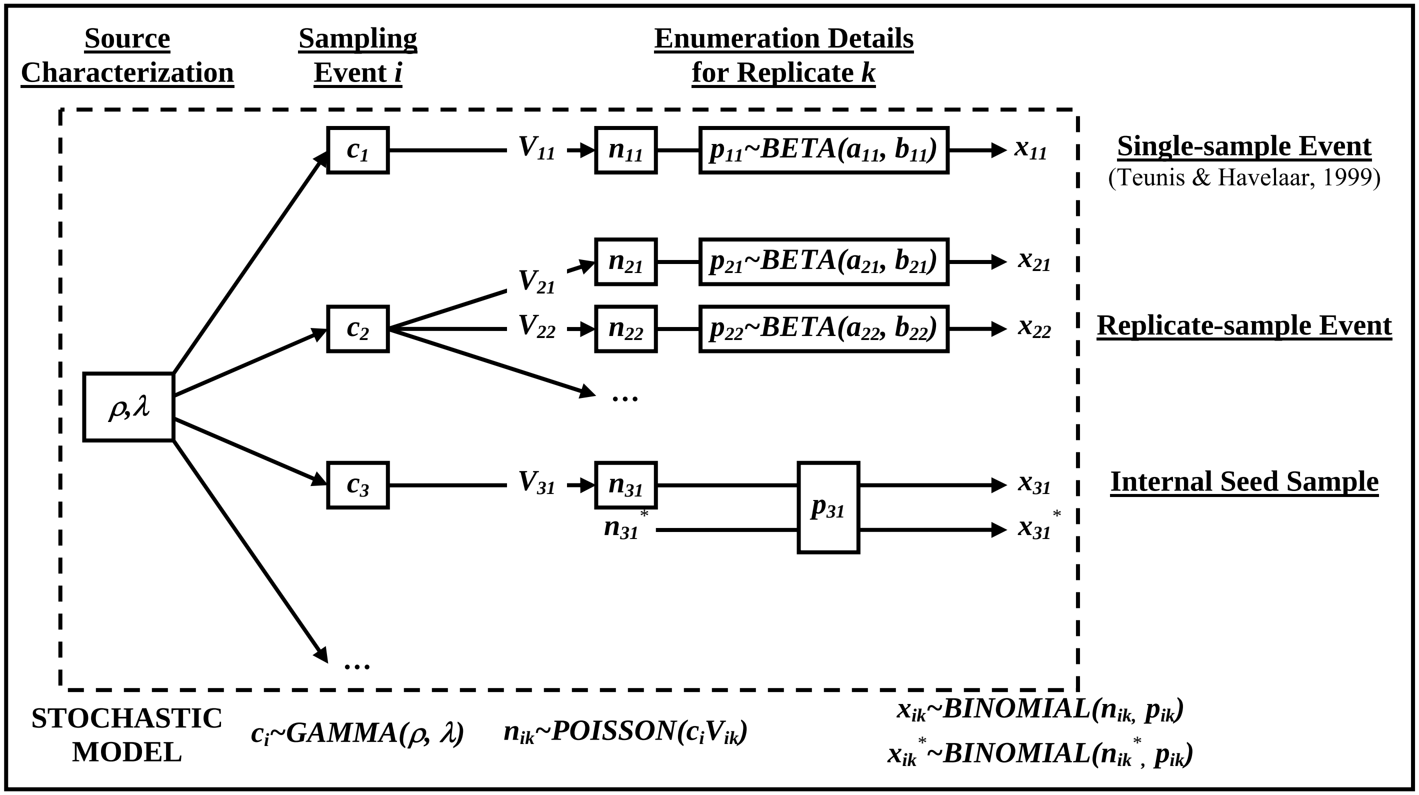

Source Characterization
Sampling Event *i*
Enumeration Details for Replicate *k*
c_1
V_{11}
n_{11}
p_{11}~BETA(a_{11}, b_{11})
x_{11}
Single-sample Event
(Teunis & Havelaar, 1999)
V_{21}
n_{21}
p_{21}~BETA(a_{21}, b_{21})
x_{21}
c_2
V_{22}
n_{22}
p_{22}~BETA(a_{22}, b_{22})
x_{22}
Replicate-sample Event
…
ρ,λ
c_3
V_{31}
n_{31}
n_{31}^*
p_{31}
x_{31}
x_{31}^*
Internal Seed Sample
…
STOCHASTIC MODEL
c_i~GAMMA(ρ, λ)
n_{ik}~POISSON(c_iV_{ik})
x_{ik}~BINOMIAL(n_{ik}, p_{ik})
x_{ik}^*~BINOMIAL(n_{ik}^*, p_{ik})

Figure B.3: Probabilistic models for random errors in temporally distributed concentration data

(Schmidt and Emelko, 2010)

Enumeration data with beta-distributed and internal seed recovery can be modelled together as shown because the models are identical except for the available information about analytical recovery. There can be any number of replicates within each of the *i* sampling events, and any or all of these may have associated internal seed recovery data. The negative binomial temporal concentration variability model is not shown in this figure.

Joint Distribution

$$f\left(\{c_i\},\{n_{ik}\},\{x_{ik}\},\{p_{ik}\}\mid \rho,\lambda,\{V_{ik}\},\{a_{ik}\},\{b_{ik}\}\right)=$$

$$\prod_{i=1}^{r}\left[\frac{c_i^{\rho-1}e^{-c_i/\lambda}}{\lambda^{\rho}\Gamma(\rho)}\right]\prod_{k=1}^{r_i}\left[\frac{e^{-c_iV_{ik}}\left(c_iV_{ik}\right)^{n_{ik}}}{x_{ik}!\left(n_{ik}-x_{ik}\right)!}p_{ik}^{x_{ik}}\left(1-p_{ik}\right)^{n_{ik}-x_{ik}}\right]\left[\frac{\Gamma\left(a_{ik}+b_{ik}\right)}{\Gamma\left(a_{ik}\right)\Gamma\left(b_{ik}\right)}p_{ik}^{a_{ik}-1}\left(1-p_{ik}\right)^{b_{ik}-1}\right]$$

Joint Posterior Distribution

$$f\left(\rho,\lambda,\{c_i\}\{n_{ik}\},\{p_{ik}\}\mid\{x_{ik}\},\{V_{ik}\},\{a_{ik}\},\{b_{ik}\}\right)\propto g\left(\rho,\lambda\right)\cdot$$

$$\prod_{i=1}^{r}\left[\frac{c_i^{\rho-1}e^{-c_i/\lambda}}{\lambda^{\rho}\Gamma(\rho)}\right]\prod_{k=1}^{r_i}\left[\frac{e^{-c_iV_{ik}}\left(c_iV_{ik}\right)^{n_{ik}}}{x_{ik}!\left(n_{ik}-x_{ik}\right)!}p_{ik}^{x_{ik}}\left(1-p_{ik}\right)^{n_{ik}-x_{ik}}\right]\left[\frac{\Gamma\left(a_{ik}+b_{ik}\right)}{\Gamma\left(a_{ik}\right)\Gamma\left(b_{ik}\right)}p_{ik}^{a_{ik}-1}\left(1-p_{ik}\right)^{b_{ik}-1}\right]$$

Conditional Posterior Distributions

$$Dfc\left(\rho\right)\propto g\left(\rho\mid\lambda\right)\cdot\left(\Gamma\left(\rho\right)\right)^{-r}\left(\frac{1}{\lambda}\prod_{i=1}^{r}c_i\right)^{\rho}$$

$$Dfc\left(\lambda\right)\propto g\left(\lambda\mid\rho\right)\cdot\left(\frac{1}{\lambda}\right)^{r\rho}e^{-\left(\sum c_i/\lambda\right)}$$

$$Dfc\left(c_i\right)\propto c_i^{\sum n_{ik}+\rho-1}e^{-c_i\left(\sum V_{ik}+1/\lambda\right)}\rightarrow c_i\left(\sum_{k=1}^{r_i}V_{ik}+\frac{1}{\lambda}\right)\sim GAMMA\left(\sum_{k=1}^{r_i}n_{ik}+\rho,1\right)$$

$$Dfc\left(n_{ik}\right)\propto\frac{\left(c_iV_{ik}\right)^{n_{ik}}}{\left(n_{ik}-x_{ik}\right)!}\left(1-p_{ik}\right)^{n_{ik}}\rightarrow n_{ik}-x_{ik}\sim POISSON\left(c_iV_{ik}\left(1-p_{ik}\right)\right)$$

$$Dfc\left(p_{ik}\right)\propto p_{ik}^{x_{ik}+a_{ik}-1}\left(1-p_{ik}\right)^{n_{ik}-x_{ik}+b_{ik}-1}\rightarrow p_{ik}\sim BETA\left(x_{ik}+a_{ik},n_{ik}-x_{ik}+b_{ik}\right)$$

Notes

This model is generalized so that each of the *i* sampling events may have any number of replicate enumeration data with potentially sample-specific volume and recovery distribution parameters. An acceptance rejection algorithm to sample (approximately) from the conditional posterior of ρ is addressed in Section 6.3.1. If $g(\lambda|\rho)$ is either an improper semi-infinite uniform prior or a gamma distribution, then the conditional posterior of λ^{-1} is gamma-distributed.

B.5.2 Internal Seed Temporal Concentration Variability Model

Definitions

$c_i \sim GAMMA(\rho, \lambda)$ particle concentration during the ith sampling event

$n_{ik} \sim POISSON(c_i V_{ik})$ number of particles in the kth replicate of the ith sampling event

$x_{ik} \sim BINOMIAL(n_{ik}, p_{ik})$ number of particles observed in the kth replicate of the ith sampling event

$x_{ik}^* \sim BINOMIAL(n_{ik}^*, p_{ik})$ number of seeded particles observed in the kth replicate of the ith sampling event

ρ,λ shape and scale parameters of the temporal concentration variability distribution

V_{ik} the sample volume of the kth replicate of the ith sampling event

p_{ik} analytical recovery of the kth replicate of the ith sampling event

r number of sampling events

r_i number of replicate samples in the ith sampling event

Joint Distribution

$$f\left(\{c_i\},\{n_{ik}\},\{x_{ik}\},\left\{x_{ik}^*\right\} \mid \rho,\lambda,\{V_{ik}\},\{p_{ik}\},\left\{n_{ik}^*\right\}\right)=$$

$$\prod_{i=1}^{r}\left[\frac{c_i^{\rho-1}e^{-c_i/\lambda}}{\lambda^{\rho}\Gamma(\rho)}\right]\prod_{k=1}^{r_i}\left[\frac{e^{-c_iV_{ik}}\left(c_iV_{ik}\right)^{n_{ik}}}{x_{ik}!\left(n_{ik}-x_{ik}\right)!}p_{ik}^{x_{ik}}\left(1-p_{ik}\right)^{n_{ik}-x_{ik}}\right]\left[\frac{n_{ik}^*!}{x_{ik}^*!\left(n_{ik}^*-x_{ik}^*\right)!}p_{ik}^{x_{ik}^*}\left(1-p_{ik}\right)^{n_{ik}^*-x_{ik}^*}\right]$$

Joint Posterior Distribution

$$f\left(\rho,\lambda,\{c_i\},\{n_{ik}\},\{p_{ik}\} \mid \{x_{ik}\},\left\{x_{ik}^*\right\},\{V_{ik}\},\left\{n_{ik}^*\right\}\right)\propto g\left(\rho,\lambda,\{p_{ik}\}\right)\cdot$$

$$\prod_{i=1}^{r}\left[\frac{c_i^{\rho-1}e^{-c_i/\lambda}}{\lambda^{\rho}\Gamma(\rho)}\right]\prod_{k=1}^{r_i}\left[\frac{e^{-c_iV_{ik}}\left(c_iV_{ik}\right)^{n_{ik}}}{x_{ik}!\left(n_{ik}-x_{ik}\right)!}p_{ik}^{x_{ik}}\left(1-p_{ik}\right)^{n_{ik}-x_{ik}}\right]\left[\frac{n_{ik}^*!}{x_{ik}^*!\left(n_{ik}^*-x_{ik}^*\right)!}p_{ik}^{x_{ik}^*}\left(1-p_{ik}\right)^{n_{ik}^*-x_{ik}^*}\right]$$

Conditional Posterior Distributions

$$Dfc(\rho)\propto g\left(\rho \mid \lambda,\{p_{ik}\}\right)\cdot\left(\Gamma(\rho)\right)^{-r}\left(\frac{1}{\lambda}\prod_{i=1}^{r}c_i\right)^{\rho}$$

$$Dfc(\lambda)\propto g\left(\lambda \mid \rho,\{p_{ik}\}\right)\cdot\left(\frac{1}{\lambda}\right)^{r\rho}e^{-\left(\sum c_i/\lambda\right)}$$

$$Dfc(c_i)\propto c_i^{\sum n_{ik}+\rho-1}e^{-c_i\left(\sum V_{ik}+1/\lambda\right)}\rightarrow c_i\left(\sum_{k=1}^{r_i}V_{ik}+\frac{1}{\lambda}\right)\sim GAMMA\left(\sum_{k=1}^{r_i}n_{ik}+\rho,1\right)$$

$$Dfc(n_{ik})\propto\frac{\left(c_iV_{ik}\right)^{n_{ik}}}{\left(n_{ik}-x_{ik}\right)!}\left(1-p_{ik}\right)^{n_{ik}}\rightarrow n_{ik}-x_{ik}\sim POISSON\left(c_iV_{ik}\left(1-p_{ik}\right)\right)$$

$$Dfc(p_{ik})\propto g\left(p_{ik} \mid \rho,\lambda\right)\cdot p_{ik}^{x_{ik}+x_{ik}^*}\left(1-p_{ik}\right)^{n_{ik}+n_{ik}^*-x_{ik}-x_{ik}^*}$$

Notes

This model is generalized so that any sample from the same source is regarded as a replicate: the volume and number of seeded particles are potentially sample-specific. An acceptance rejection algorithm to sample (approximately) from the conditional posterior of ρ is addressed in Section 6.3.1. If $g(\lambda|\rho,\{p_{ik}\})$ is either an improper semi-infinite uniform prior or a gamma distribution, then the conditional posterior of λ^{-1} is gamma-distributed. If a beta-distributed prior (including the standard uniform distribution) is used for the probability of recovery, then its conditional posterior is beta-distributed (as discussed in Section 5.3.4). If a potentially sample-specific beta-distributed prior (p_i~Beta(a_i^*,b_i^*)) is used, then the conditional posterior for the probability of recovery is beta-distributed with the parameters ($x_i + x_i^* + a_i^*$, $n_i - x_i + n_i^* - x_i^* + b_i^*$). The standard uniform prior is the special case in which $a_i^* = 1$, $b_i^* = 1$. Accordingly, it is equivalent to incorporate internal seed recovery data and beta-distributed priors for the probability of recovery into temporal enumeration data analysis as presented in this section or as a special case of the beta-Poisson model with the sample-specific beta distribution parameters $a_i = x_i^* + a_i^*$, $b_i = n_i^* - x_i^* + b_i^*$. In this case, the beta distribution becomes an informative prior upon the probability of recovery for indigenous particles rather than a description of non-constant analytical recovery.

B.5.3 Negative Binomial Temporal Concentration Variability Model

Definitions

$c_i \sim GAMMA(\rho,\lambda)$ particle concentration during the i^{th} sampling event

$x_{ik} \sim POISSON(c_i V_{ik} p_{ik})$ number of particles observed in the k^{th} replicate of the i^{th} sampling event

$p_{ik} \sim GAMMA(\alpha_{ik},\beta_{ik})$ analytical recovery of the k^{th} replicate of the i^{th} sampling event

ρ,λ shape and scale parameters of the temporal concentration variability distribution

V_{ik} the sample volume of the k^{th} replicate of the i^{th} sampling event

α_{ik},β_{ik} gamma distribution shape and scale parameters for the k^{th} replicate, i^{th} sampling event

r number of sampling events

r_i number of replicate samples in the i^{th} sampling event

Joint Distribution

$$f(\{c_i\},\{x_{ik}\},\{p_{ik}\}|\,\rho,\lambda,\{V_{ik}\},\{\alpha_{ik}\},\{\beta_{ik}\})=$$

$$\prod_{i=1}^{r}\left[\frac{c_i^{\rho-1}e^{-c_i/\lambda}}{\lambda^{\rho}\Gamma(\rho)}\right]\prod_{k=1}^{r_i}\left[\frac{e^{-c_i V_{ik} p_{ik}}(c_i V_{ik} p_{ik})^{x_{ik}}}{x_{ik}!}\right]\left[\frac{1}{\beta_{ik}^{\alpha_{ik}}\Gamma(\alpha_{ik})}p_{ik}^{\alpha_{ik}-1}e^{-p_{ik}/\beta_{ik}}\right]$$

Joint Posterior Distribution

$$f(\rho,\lambda,\{c_i\},\{p_{ik}\}|\,\{x_{ik}\},\{V_{ik}\},\{\alpha_{ik}\},\{\beta_{ik}\})\propto g(\rho,\lambda)\cdot$$

$$\prod_{i=1}^{r}\left[\frac{c_i^{\rho-1}e^{-c_i/\lambda}}{\lambda^{\rho}\Gamma(\rho)}\right]\prod_{k=1}^{r_i}\left[\frac{e^{-c_i V_{ik} p_{ik}}(c_i V_{ik} p_{ik})^{x_{ik}}}{x_{ik}!}\right]\left[\frac{1}{\beta_{ik}^{\alpha_{ik}}\Gamma(\alpha_{ik})}p_{ik}^{\alpha_{ik}-1}e^{-p_{ik}/\beta_{ik}}\right]$$

Conditional Posterior Distributions

$$Dfc(\rho) \propto g(\rho \mid \lambda) \cdot (\Gamma(\rho))^{-r} \left(\frac{1}{\lambda} \prod_{i=1}^{r} c_i \right)^{\rho}$$

$$Dfc(\lambda) \propto g(\lambda \mid \rho) \cdot \left(\frac{1}{\lambda} \right)^{r\rho} e^{-\left(\sum c_i / \lambda\right)}$$

$$Dfc(c_i) \propto c_i^{\sum x_{ik} + \rho - 1} e^{-c_i \left(\sum V_{ik} + 1/\lambda\right)} \rightarrow c_i \left(\sum_{k=1}^{r_i} V_{ik} + \frac{1}{\lambda} \right) \sim GAMMA\left(\sum_{k=1}^{r_i} x_{ik} + \rho, 1 \right)$$

$$Dfc(p_{ik}) \propto p_{ik}^{x_{ik} + \alpha_{ik} - 1} e^{-p_{ik}\left(c_i V_{ik} + 1/\beta_{ik}\right)} \rightarrow \frac{p_{ik}}{\beta_{ik}} \left(c_i V_{ik} \beta_{ik} + 1 \right) \sim GAMMA\left(x_{ik} + \alpha_{ik}, 1 \right)$$

Notes

This model is generalized so that each of the *i* sampling events may have any number of replicate enumeration data with potentially sample-specific volume and recovery distribution parameters. An acceptance rejection algorithm to sample (approximately) from the conditional posterior of ρ is addressed in Section 6.3.1. If $g(\lambda|\rho)$ is either an improper semi-infinite uniform prior or a gamma distribution, then the conditional posterior of λ^{-1} is gamma-distributed.

Appendix C
Sample Visual Basic Code for Calculation of Posterior Concentration Distributions

This appendix contains sample Visual Basic^{TM} code for various algorithms associated with the computation of posterior concentration distributions. In particular, it provides pseudo-code for numerical integration of the integral in Equation 5.16 (Appendix C.1), and for the Gibbs sampling algorithm that is presented in Figure 5.4 (Appendix C.2).

C.1 SAMPLE CODE FOR NUMERICAL INTEGRATION OF THE BETA-POISSON ENUMERATION MODEL

Numerical integration of Equation 5.16 (shown below) to obtain posterior probability densities for concentration in the beta-Poisson enumeration model requires an algorithm to approximate the integral within the equation (Appendix C.1.1) as well as numerical integration of the right hand side for c > 0 (Appendix C.1.2) to approximate the integrating factor (*i.e.* the constant that is needed to make the proportionality an equation.

$$f\left(c \mid \left\{x_i\right\}\right) \propto g(c) \cdot \prod_{i=1}^{r} c^{x_i} \cdot \int_0^1 e^{-cV_i p}\, p^{x_i+a-1}(1-p)^{b-1}\, dp$$

C.1.1 Numerical Integration of Beta-Poisson Likelihoods

Numerical integration of the integral within Equation 5.16 was addressed in Schmidt *et al.* (2010b), but the result was inaccurate if the integrand had a narrow peak. A better algorithm must integrate with better resolution across the peak of the integrand. The code that is shown herein integrates using 1000 equal segments of the interval 0 < p < 1 and would need to be revised if the integrand has a narrow peak.

Location of maximum

To locate the peak of the integrand, the mode of the function $f(p)$ must be found.

$$f(p) = e^{-cVp}\, p^{x+a-1}(1-p)^{b-1}$$

$$\ln(f(p)) = -cVp + (x+a-1)\cdot \ln(p) + (b-1)\cdot \ln(1-p) \qquad p \neq 0, p \neq 1$$

$$\frac{d\ln(f(p))}{dp} = -cV + \frac{(x+a-1)}{p} - \frac{(b-1)}{(1-p)} = 0$$

$$-cVp(1-p)+(x+a-1)(1-p)-(b-1)p=0$$

$$cVp^2-(cV+x+a+b-2)p+(x+a-1)=0$$

The roots of this equation can be evaluated using the quadratic formula. The maximum will occur at one of these points within the interval $0 < p < 1$ (because $f(p) = 0$ for $p = 0$ or $p = 1$ but is positive for all values in between).

Sample Code

The function `LNINTEGRAL` calculates the approximate natural logarithm of the integral within Equation 5.16. The integration could be refined by evaluating more than 1000 segments or by using narrower segments near the mode.

```
‘    Externally defined input variables in function LNINTEGRAL ()
     Dim sgCV As Single      ‘    Product of conc. value and sample volume
     Dim inX as Integer      ‘    Count value
     Dim sgA as Single       ‘    Beta distribution recovery parameter
     Dim sgB as Single       ‘    Beta distribution recovery parameter

Private Function LNINTEGRAL(sgCV, inX, sgA, sgB) As Double

‘    Declare local variables
     Dim inI As Integer      ‘    Index variable
     Dim sgP As Single       ‘    Recovery constant value
     Dim dbLnInt() As Double ‘    Natural logarithm of integrand
     Dim dbLnMax As Double   ‘    Maximum of ln(integrand)

‘    Locate maximum wrt sgP (code not shown) and calculate ln of maximum
     dbLnMax = -sgCV * sgP + (inX+sgA–1) * Log(sgP) + (sgB–1) * Log(1-sgP)

‘    Evaluate ln(integrand) at midpoint of each of 1000 segments
     ReDim dbLnInt(1 to 1000)
     For inI = 1 to 1000
          sgP = (inI – 0.5) / 1000
          ‘    Subtract dbLnMax from each value and add back later to
          ‘    prevent overflow of summation
          dbLnInt(inI) = -sgCV * sgP + (inX + sgA - 1) * Log(sgP) + _
               (sgB - 1) * Log(1 - sgP) - dbLnMax
    Next

'   Calculate integral
    For inI = 1 To 1000
        LNINTEGRAL = LNINTEGRAL + Exp(dbLnInt(inI) - dbLnMax) * 0.001
    Next
    LNINTEGRAL = Log(LNINTEGRAL) + dbLnMax

End Function
```

C.1.2 Evaluation of the Integrating Factor in Equation 5.16

The integrating factor that is required to convert Equation 5.16 from a proportionality to a probability density function must be evaluated by numerical integration. In general, the procedure consists of locating and evaluating the mode, defining an interval of concentration values outside which the posterior density is trivially small in comparison to the mode, and evaluating the integrating factor across this interval. Once this analysis has been completed, the posterior probability density for any concentration value can be calculated using the approximated integrating factor and the formula in the right hand side of Equation 5.16. The code shown below demonstrates how to evaluate the integral across a defined interval after the maximum value has been evaluated. It uses the `LNINTEGRAL` function discussed in Appendix C.1.1 and a function `LNPRIOR` that returns the logarithm of the prior *g*(*c*) (which has a value of zero if a uniform prior density of 1 is used).

Sample Code

```
‘    Declare input variables
     Dim inR as Integer      ‘   Number of replicate enumeration data
     Dim inX() as Integer    ‘   Array of sample counts (1 to inR)
     Dim sgV() as Single     ‘   Array of sample volumes (1 to inR)
     Dim sgA as Single       ‘   Beta distribution recovery parameter
     Dim sgB as Single       ‘   Beta distribution recovery parameter

‘    Declare other variables
     Dim dbLnMax as Double   ‘   Maximum posterior density value
     Dim sgConc as Single    ‘   Concentration value
     Dim inI as Integer      ‘   Index variable (sample number)
     Dim sgCV as Single      ‘   Product of conc. value and sample volume
     Dim lnNumStep as Long   ‘   # of segments in numerical integration
     Dim sgLowC as Single    ‘   Lower boundary of concentration interval
     Dim sgHighC as Single   ‘   Upper boundary of concentration interval
     Dim dbK as Double       ‘   Integrating factor in post. distribution
     Dim lnJ as Long         ‘   Index variable (segment number)
     Dim dbLnF               ‘   Natural logarithm of posterior density

‘    Input data (code not shown)

‘    Locate maximum wrt sgConc (code not shown) and calculate ln of maximum
     dbLnMax = LNPRIOR(sgConc)
     For inI = 1 to inR
          sgCV = sgConc * sgV(inI)
          dbLnMax = dbLNMax + inX(inI) * Log(sgConc) + _
               LNINTEGRAL(sgCV, inX(inI), sgA, sgB)
     Next

‘    Evaluate natural logarithm of posterior across lnNumStep segments of
‘    the interval sgLowC < sgConc < sgHighC
     dbK = 0
     For lnJ = 1 to lnNumStep
          sgConc = sgLowC + (lnJ - 0.5) / lnNumStep * (sgHighC – sgLowC)
          ‘    Evaluate ln of posterior density at midpoint of each segment
          ‘    Subtract dbLnMax from each value to avoid overflows in sum
          ‘    Add dbLnMax after sum computed
          dbLnF = LNPRIOR(sgConc)
          For inI = 1 to inR
               sgCV = sgConc * sgV(inI)
               dbLnF = dbLnF + inX(inI) * Log(sgConc) + _
                    LNINTEGRAL(sgCV, inX(inI), sgA, sgB) - dbLnMax
          Next
          ‘    Add segment area to integral
          dbK = dbK  + Exp(dbLnF) * (sgHighC – sgLowC) / lnNumStep
     Next
     ‘    Add dbLnMax back into integral
     dbK = dbK * Exp(dbLnMax)
```

C.2 Sample Gibbs Sampling Code

The following code provides a framework for Gibbs sampling associated with the beta-Poisson enumeration model presented in Figure 5.4. The code is written using the Visual Basic Editor in Microsoft Excel. “The user must (1) modify the code to input data, (2) provide functions for random number generation, and (3) program data output and analysis (*e.g.* calculation of summary statistics such as the mean, mode, standard deviation, and 95% credible interval)” (Schmidt *et al.*, 2010b). The code that is provided has been modified slightly from Schmidt *et al.*, 2010b to reflect the changed sequence of conditional posterior distributions and initial parameter values.

```
‘    Declare input variables
     Dim inR as Integer     ‘    Number of replicate enumeration data
     Dim inX() as Integer   ‘    Array of sample counts (1 to inR)
     Dim sgV() as Single    ‘    Array of sample volumes (1 to inR)
     Dim sgA as Single      ‘    Beta distribution recovery parameter
     Dim sgB as Single      ‘    Beta distribution recovery parameter
     Dim lnNumBurn as Long  ‘    Number of burn-in iterations
     Dim lnNumSave as Long  ‘    Number of iterations to save in posterior

‘    Declare temporary/output variables
     Dim inI as integer     ‘    Sample index
     Dim sgP() as Single    ‘    Array of unknown recoveries
     Dim inN() as Integer   ‘    Array of unknown true counts
     Dim sgConc As Single   ‘    Temporary concentration value
     Dim lnSumX as Long     ‘    Sum of counts
     Dim sgSumVP as Single  ‘    Sum of sample volumes
     Dim sgC() as Single    ‘    Sequence of posterior conc. values
     Dim lnTrial as Long    ‘    Gibbs sampling iteration index
     Dim lnSumN as Long     ‘    Sum of unknown true counts
     Dim sgCV as Single     ‘    Product of concentration and Sum(Volume)
     Dim sgLamda as Single  ‘    Temporary Poisson parameter

‘    Input data (code not shown)

‘    Assign initial values for parameters
     ReDim sgP(1 to inR)
     ReDim inN(1 to inR)
     lnSumX = 0
     sgSumVP = 0
     For inI = 1 to inR
          ‘    Assign initial values of Pi
          sgP(inI) = sgA/(sgA + sgB)
          lnSumX = lnSumX + inX(inI)
          sgSumVP = sgSumVP + sgV(inI)*sgP(inI)
     Next
     ‘    Assign non-zero initial value of C
     If lnSumX > 0 Then
          sgConc = lnSumX/lnSumVP
     Else
          sgConc = 0.5/lnSumVP
     End If

‘    Run Gibbs sampling
     ReDim sgC(1 to lnNumSave)
     For lnTrial = 1 to lnNumBurn + lnNumSave
          ‘    Generate new values of Ni from Dfc(Ni)
          lnSumN = 0
          For inI = 1 to inR
               inN(inI) = POISSON(sgLamda) + inX(inI)
               lnSumN = lnSumN + inN(inI)
          Next
          ‘    Generate new value of C from Dfc(C)
          sgCV = GAMMA(lnSumN + 1)
          sgConc = sgCV / sgSumV
          ‘    Store concentration values after burn-in
          If lnTrial > lnNumBurn Then
               sgC(lnTrial – lnNumBurn) = sgConc
          End If
          ‘    Generate new values of Pi from Dfc(Pi)
          For inI = 1 to inR
               sgP(inI) = BETA(inX(inI) + sgA, inN(inI) – inX(inI) +sgB)
               sgLamda = sgConc * sgV(ini) * (1 – sgP(inI))
          Next
     Next
```

Appendix D
Case Study Data

This appendix contains case study datasets that are used in this thesis but that are not presented in the main body of the thesis or readily available in other publications. The data that are presented include (1) the ongoing precision and recovery data discussed in Section 4.4.2, (2) the manual hemocytometer microsphere counts discussed in Section 5.2.2, (3) the recovery and stable filter operation log-removal data used in Section 5.4.4, and (4) the simulated temporally distributed enumeration data used in Sections 6.3.2 and 6.5.4.

D.1 Ongoing Precision and Recovery Data

Table D.1 presents the ongoing precision and recovery data that were discussed in Section 4.4.2. These data were obtained over several years by an anonymous laboratory that is approved for use of Method 1623 (USEPA, 2005) to enumerate *Cryptosporidium* oocysts and *Giardia* cysts in water. To protect confidentiality, the data are presented only as seed doses and obtained counts (without dates or seed dose lot numbers). The index numbers correspond to the chronological order of the data.

Table D.1: Ongoing precision and recovery dataset

	Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Crypto.	**Count**	34	48	43	33	21	47	48	53	53	50	46	34	35	41	25	53	40	40	21	55	34	57	52	35	53
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
Giardia	**Count**	53	46	67	47	43	54	48	59	70	61	48	53	58	42	48	48	55	40	33	50	50	52	44	36	55
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	98	98	98
	Index	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Crypto.	**Count**	20	27	47	55	40	42	53	38	54	32	28	50	44	37	31	41	30	34	35	23	45	44	31	35	75
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	98	98	98	98	98	98	98	98	98	98
Giardia	**Count**	37	23	51	52	50	56	61	39	40	33	46	51	32	28	34	27	36	40	48	37	30	51	31	56	61
	Dose	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
	Index	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Crypto.	**Count**	36	52	46	25	54	26	41	27	32	25	20	34	49	40	41	32	52	42	28	37	50	33	39	45	48
	Dose	98	98	98	98	98	98	98	98	98	98	98	99	99	99	99	99	99	99	99	99	99	100	100	100	100
Giardia	**Count**	26	66	38	50	55	37	45	51	41	41	44	57	37	39	44	41	60	45	61	41	52	49	51	43	45
	Dose	98	98	98	98	98	98	98	98	98	98	98	99	99	99	99	99	99	99	99	99	99	100	100	100	100
	Index	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
Crypto.	**Count**	33	30	39	33	46	40	41	38	41	54	36	43	51	53	57	53	36	36	70	54	34	44	55	42	44
	Dose	100	100	100	100	100	100	100	100	100	100	100	99	99	99	99	99	99	99	99	99	99	99	99	99	99
Giardia	**Count**	36	35	56	41	48	43	59	45	53	52	37	46	58	53	52	28	51	22	61	38	36	42	35	39	45
	Dose	100	100	100	100	100	100	100	100	100	100	100	99	99	99	99	99	99	99	99	99	99	99	99	99	99
	Index	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125
Crypto.	**Count**	42	25	54	34	39	43	27	9	47	50	49	31	33	30	36	11	36	19	48	26	46	52	43	44	44
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
Giardia	**Count**	39	44	43	81	43	55	51	61	44	58	52	42	53	55	43	51	42	54	39	56	39	51	51	60	40
	Dose	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Index	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150
Crypto.	**Count**	49	29	41	57	44	45	38	55	39	32	53	44	40	46	47	44	48	41	46	58	45	50	52	49	49
	Dose	99	99	99	100	100	100	100	100	100	100	100	100	100	100	100	100	99	99	99	99	99	99	99	99	99
Giardia	**Count**	30	33	38	47	48	49	22	59	37	38	38	65	50	47	44	47	49	51	45	52	50	51	34	33	48
	Dose	100	100	100	99	99	99	99	99	99	99	99	99	99	99	99	99	98	98	98	98	98	98	98	98	98

Table D.1: Ongoing precision and recovery dataset (continued)

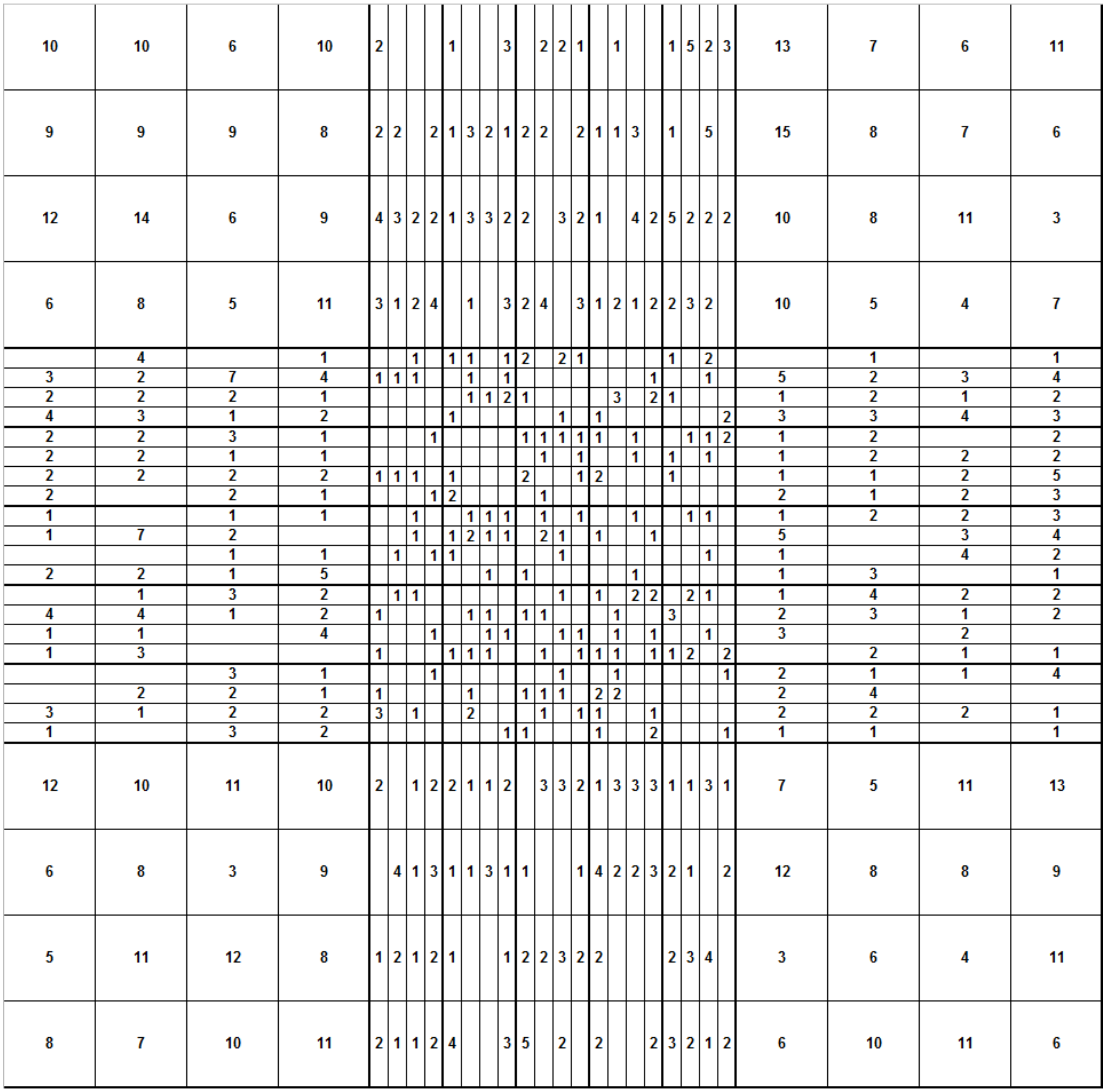
	Index	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
Crypto.	**Count**	43	39	56	38	46	55	60	48	47	49	46	24	40	44	29	41	45	39	43	27	40	42	42	38	35
	Dose	99	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	98	98	98	98	98	98	98	98	98
Giardia	**Count**	40	60	62	35	60	64	60	51	49	46	50	33	37	42	45	38	30	45	42	50	46	26	41	45	42
	Dose	98	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
	Index	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
Crypto.	**Count**	35	40	40	33	41	38	45	66	47	52	58	53	50	58	15	61	54	47	55	45	49	48	50	55	42
	Dose	98	98	98	98	98	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	100	100	100	100	100
Giardia	**Count**	36	38	34	40	39	28	34	30	40	22	48	45	45	41	38	38	41	39	30	33	50	28	39	59	37
	Dose	99	99	99	99	99	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	99	99	99	99	99
	Index	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225
Crypto.	**Count**	54	64	52	54	53	63	55	48	59	55	51	37	51	40	45	48	50	51	36	38	53	38	33	33	39
	Dose	100	100	100	100	100	100	100	100	100	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
Giardia	**Count**	53	40	42	37	44	46	26	46	37	46	47	52	49	47	31	35	43	48	38	48	45	42	40	44	36
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
	Index	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250
Crypto.	**Count**	36	32	34	38	36	41	37	46	30	50	40	44	40	41	32	51	54	46	46	47	33	43	51	38	38
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
Giardia	**Count**	44	36	27	24	27	26	30	55	31	36	40	20	31	8	66	33	64	40	40	50	35	37	23	44	48
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	100	100	100	100	100
	Index	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275
Crypto.	**Count**	42	43	52	38	56	52	56	48	36	56	49	51	25	45	46	41	36	41	50	35	49	47	46	37	33
	Dose	99	99	99	99	99	99	99	99	99	99	99	100	100	100	100	100	100	100	100	100	100	100	100	100	99
Giardia	**Count**	52	48	56	30	61	52	48	43	51	44	46	32	36	45	50	33	53	45	34	28	43	43	42	37	44
	Dose	100	100	100	100	100	98	98	98	98	98	98	99	99	99	99	99	99	99	99	99	99	99	99	99	98
	Index	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300
Crypto.	**Count**	36	42	48	47	38	42	30	54	49	30	63	56	45	36	50	33	58	54	40	41	62	49	58	48	64
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
Giardia	**Count**	42	20	26	33	46	59	36	51	66	28	36	44	41	30	45	33	38	47	30	34	54	38	26	32	60
	Dose	98	98	98	98	98	98	98	98	98	98	98	98	98	99	99	99	99	99	99	99	99	99	99	99	99

Table D.1: Ongoing precision and recovery dataset (continued)

	Index	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325
Crypto.	**Count**	45	53	49	52	58	69	64	49	47	45	53	57	35	55	57	37	67	63	25	38	59	35	54	61	47
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	98	98	98	98	98	98	98	98	98	100	100	100	100
Giardia	**Count**	45	49	31	28	44	57	57	36	31	41	41	46	53	45	49	25	63	40	37	49	37	35	50	38	24
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
	Index	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350
Crypto.	**Count**	56	38	48	47	46	36	53	65	56	49	46	29	55	27	47	54	41	41	48	43	46	43	58	57	52
	Dose	100	100	100	100	100	100	100	100	100	100	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Giardia	**Count**	42	37	54	40	43	42	43	46	56	56	30	37	40	26	41	33	30	54	49	31	46	42	62	52	49
	Dose	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	98	98	98	98	98	98	99	99
	Index	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375
Crypto.	**Count**	63	44	63	58	53	48	52	57	47	48	60	63	44	60	47	49	43	49	56	48	54	52	49	30	28
	Dose	98	98	98	98	98	98	99	99	99	99	99	99	99	99	99	99	100	100	100	100	100	100	100	100	100
Giardia	**Count**	47	37	52	55	49	62	45	48	46	37	46	48	44	59	48	55	43	47	46	24	50	37	44	50	47
	Dose	99	99	99	99	99	99	98	98	98	98	98	98	98	99	99	99	99	99	99	99	99	99	100	100	100
	Index	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400
Crypto.	**Count**	52	28	16	59	18	50	22	50	31	35	33	36	28	47	46	61	59	52	58	50	42	52	51	48	67
	Dose	100	100	100	100	100	100	100	100	99	99	99	99	99	99	99	99	99	100	100	100	100	100	100	100	100
Giardia	**Count**	47	47	26	57	51	68	42	57	64	56	48	37	47	42	42	51	57	48	65	55	32	50	45	43	43
	Dose	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Index	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425
Crypto.	**Count**	52	46	36	43	53	31	13	66	54	54	53	58	60	59	52	32	55	50	31	22	41	45	41	59	60
	Dose	101	101	101	101	101	101	99	99	99	99	99	99	100	100	100	100	100	100	101	101	101	101	101	101	101
Giardia	**Count**	58	48	49	49	47	50	54	56	61	43	45	42	34	36	39	24	18	34	35	40	31	53	58	53	53
	Dose	100	100	100	100	100	100	101	101	101	101	101	101	101	101	101	101	101	101	100	100	100	100	100	100	100
	Index	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444						
Crypto.	**Count**	52	51	53	49	45	51	45	52	50	43	53	50	37	33	56	46	61	58	50						
	Dose	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	99	99						
Giardia	**Count**	38	62	46	48	65	63	41	59	42	39	46	55	36	47	52	51	54	51	26						
	Dose	101	101	101	101	101	100	100	100	100	100	100	100	100	100	100	100	100	99	99						

D.2 MICROSPHERE COUNTS ON A PETROFF-HAUSSER COUNTING CHAMBER

Figure D.1 presents the manual quadrat counts of polystyrene microspheres upon the improved Neubauer grid of a Petroff-Hausser counting chamber as discussed in Section 5.2.2.



10	10	6	10	2				1			3		2	2	1		1			1	5	2	3	13	7	6	11
9	9	9	8	2	2		2	1	3	2	1	2	2		2	1	1	3		1		5		15	8	7	6
12	14	6	9	4	3	2	2	1	3	3	2	2		3	2	1		4	2	5	2	2	2	10	8	11	3
6	8	5	11	3	1	2	4		1		3	2	4		3	1	2	1	2	2	3	2		10	5	4	7
	4		1			1		1	1		1	2		2	1					1		2			1		1
3	2	7	4	1	1	1			1		1								1			1		5	2	3	4
2	2	2	1						1	1	2	1					3		2	1				1	2	1	2
4	3	1	2					1						1		1							2	3	3	4	3
2	2	3	1				1					1	1	1	1	1		1			1	1	2	1	2		2
2	2	1	1										1		1			1		1		1		1	2	2	2
2	2	2	2	1	1	1		1				2			1	2				1				1	1	2	5
2		2	1				1	2					1											2	1	2	3
1		1	1			1			1	1	1		1		1			1			1	1		1	2	2	3
1	7	2				1		1	2	1	1		2	1		1			1					5		3	4
		1	1		1		1	1						1								1		1		4	2
2	2	1	5							1		1						1						1	3		1
	1	3	2		1	1								1		1		2	2		2	1		1	4	2	2
4	4	1	2	1					1	1		1	1				1			3				2	3	1	2
1	1		4				1			1	1			1	1		1		1			1		3		2	
1	3			1				1	1	1			1		1	1	1		1	1	2		2		2	1	1
		3	1				1							1			1						1	2	1	1	4
	2	2	1	1					1			1	1	1		2	2							2	4		
3	1	2	2	3		1			2				1		1	1			1					2	2	2	1
1		3	2								1	1				1			2				1	1	1		1
12	10	11	10	2		1	2	2	1	1	2		3	3	2	1	3	3	3	1	1	3	1	7	5	11	13
6	8	3	9		4	1	3	1	1	3	1	1			1	4	2	2	3	2	1		2	12	8	8	9
5	11	12	8	1	2	1	2	1			1	2	2	3	2	2				2	3	4		3	6	4	11
8	7	10	11	2	1	1	2	4			3	5		2		2			2	3	2	1	2	6	10	11	6

Figure D.1: Quadrat counts of microspheres on a Petroff-Hausser counting chamber

The figure shows the approximate layout of the 3 mm × 3 mm improved Neubauer grid.

D.3 Example log-Reduction Data

The data summarized in this section are used as a case study in Section 5.4.4. Table D.2 presents the relevant recovery information. The seed and count data shown in this table are hypothetical values that are used to enable the analyses presented in Section 5.4.4.

Table D.2: ***Cryptosporidium*** **recovery data in filter influent and filter effluent***

Location	**Volume**	**Seeded Conc.**	**Measured Conc.**	**Recovery**	**Seed**†	**Count**†
	(mL)	(oocysts/L)	(oocysts/L)	(%)	(-)	(-)
Filter Influent	10	2.0E+05	5.5E+04	28	2000	550
	10	2.0E+05	6.1E+04	31	2000	610
	10	2.0E+05	4.8E+04	24	2000	480
	10	2.0E+05	5.6E+04	28	2000	560
	10	2.0E+05	4.2E+04	21	2000	420
	10	2.0E+05	4.0E+04	20	2000	400
	10	2.0E+05	5.4E+04	27	2000	540
	10	1.0E+04	1.6E+03	16	100	16
	10	1.0E+04	2.3E+03	23	100	23
	10	1.0E+04	2.6E+03	26	100	26
	10	2.0E+05	5.7E+04	28	2000	570
	10	2.0E+05	5.8E+04	29	2000	580
	10	2.0E+05	5.2E+04	26	2000	520
	10	2.0E+05	7.8E+04	39	2000	780
	10	2.0E+05	7.2E+04	36	2000	720
	10	2.0E+05	7.5E+04	38	2000	750
	10	2.0E+05	4.6E+04	23	2000	460
	10	1.1E+04	2.7E+03	25	110	27
	10	1.1E+04	2.5E+03	23	110	25
	10	1.1E+04	4.4E+03	40	110	44
Filter Effluent	500	2.1E+02	6.4E+01	31	104	32
	500	2.1E+02	7.0E+01	34	104	35
	500	2.1E+02	6.0E+01	29	104	30
	500	2.1E+02	7.6E+01	37	104	38
	500	2.1E+02	6.4E+01	31	104	32
	500	2.1E+02	7.2E+01	35	104	36
	500	2.1E+02	5.2E+01	25	104	26
	500	2.0E+02	8.4E+01	41	102	42
	500	2.0E+02	8.8E+01	43	102	44
	500	2.0E+02	8.8E+01	43	102	44
	500	2.0E+02	5.0E+01	25	102	25
	500	2.0E+02	9.0E+01	44	102	45
	500	2.0E+02	1.1E+02	54	102	55
	500	2.0E+02	9.6E+01	47	102	48
	500	2.0E+05	5.5E+04	28	2000	550
	500	2.0E+05	6.1E+04	31	2000	610
	500	2.0E+05	4.8E+04	24	2000	480
	500	2.0E+05	5.6E+04	28	2000	560
	500	2.0E+05	4.2E+04	21	2000	420
	500	2.0E+05	4.0E+04	20	2000	400

* Modified from Table A.1 of Huck *et al.* (2001) † Hypothetical values

Table 5.4 presents the relevant log-reduction data that are available in Huck *et al.* (2001). The sample volumes and raw count data are not provided in Huck et al. (2001), so hypothetical counts and volumes were needed to facilitate the analyses presented in Section 5.4.4. These counts and volumes are summarized in Table D.3.

Table D.3: *Cryptosporidium* log-reduction data obtained during stable filter operation*

Sampling Time (min)	**Initial Volume**† (L)	**Initial Count**† (-)	**Initial Conc.** (oocysts/L)	**Final Volume**† (L)	**Final Count**† (-)	**Final Conc.** (oocysts/L)	**Log-removal** (-)
15	0.005	474	9.48×10^4	0.1	24	2.40×10^2	2.6
30	0.005	418	8.36×10^4	0.1	19	1.90×10^2	2.6
40	0.005	321	6.42×10^4	0.1	19	1.90×10^2	2.5
50	0.005	293	5.86×10^4	0.1	12	1.20×10^2	2.7

* Modified from the 7/15/98 stable filter operation experiment data – Table B.10 of Huck *et al.* (2001)

† Hypothetical values

D.4 EXAMPLE TEMPORALLY DISTRIBUTED ENUMERATION DATA

The temporally distributed enumeration data that were used in Sections 6.3.2 and 6.5.4 are shown in Table D.4. These data were simulated using the beta-Poisson temporal concentration variability model and the parameters $\rho = 0.22$, $\lambda = 0.36$, $a = 2$, and $b = 3$ (Schmidt and Emelko, 2010a).

Table D.4: Temporally distributed enumeration data (Schmidt and Emelko, 2010b)

i	V_i (L)	n_i^*	x_i	x_i^*
1	100	100	0	35
2	100	100	10	76
3	100	100	0	2
4	100	100	0	82
5	100	100	0	46
6	100	100	0	46
7	100	100	6	16
8	100	100	4	8
9	100	100	0	50
10	100	100	42	77
11	100	100	0	14
12	100	100	0	30
13	100	100	0	61
14	100	100	1	32
15	100	100	0	26
16	100	100	0	59
17	100	100	0	18
18	100	100	2	47
19	100	100	1	44
20	100	100	0	21
21	100	100	0	83
22	100	100	6	23
23	100	100	0	59
24	100	100	0	50

Appendix E
Variance Decomposition

This appendix contains the derivations of variance decomposition equations that are presented in the thesis. Variance decomposition of probabilistic recovery models (Appendix E.1), probabilistic concentration models (Appendix E.2) and a model comparing the effects of analytical recovery and infectivity upon pathogen concentration estimates (Appendix E.3) are addressed.

E.1 Variance Decomposition of Probabilistic Recovery Models

This section provides derivations in support of the variance decomposition analyses in Section 8.1. Three different models and their partial sample analysis counterparts are considered.

E.1.1 Binomial Recovery Model with Known Number of Seeded Particles

This section addresses the situation in which analytical error is modelled by a binomial distribution and analytical recovery varies randomly with mean μ_p and variance $\sigma_p{}^2$. The results of this analysis are discussed in Section 8.1.1 and are summarized by Equation 8.1.

Definitions

n	number of seeded analyte particles
p	analytical recovery of the enumeration method
x	count of analyte particles

Model

$$x \sim BINOMIAL(n, p) \qquad E[x \mid n, p] = np \qquad Var[x \mid n, p] = np(1 - p)$$

Variance Decomposition to Address Non-constant Analytical Recovery

$$E[x \mid n] = E[E[x \mid n, p]] = nE[p] = n\mu_p$$

$$Var[x \mid n] = E[Var[x \mid n, p]] + Var[E[x \mid n, p]] = nE[p] - nE\left[p^2\right] + n^2 Var[p] = n\mu_p - n\left(\mu_p^2 + \sigma_p^2\right) + n^2\sigma_p^2$$

$E\left[\frac{x}{n} \mid n\right] = \mu_p \quad \therefore$ x/n is an unbiased estimate of mean analytical recovery

$$Var\left[\frac{x}{n} \mid n\right] = \frac{\mu_p}{n} - \frac{\left(\mu_p^2 + \sigma_p^2\right)}{n} + \sigma_p^2$$

Beta-distributed Analytical Recovery

$$\mu_p = \frac{a}{a+b} \qquad 1-\mu_p = \frac{b}{a+b} \qquad \sigma_p^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

$$Var\left[\frac{x}{n}\,|\,n\right] = \frac{\mu_p}{n} - \frac{\left(\mu_p^2+\sigma_p^2\right)}{n} + \sigma_p^2 = \frac{\mu_p\left(1-\mu_p\right)-\sigma_p^2}{n} + \sigma_p^2$$

$$\mu_p\left(1-\mu_p\right)-\sigma_p^2 = \frac{ab}{(a+b)^2} - \frac{ab}{(a+b)^2(a+b+1)} = \frac{ab}{(a+b)^2(a+b+1)}(a+b+1-1) = \frac{ab}{(a+b)(a+b+1)}$$

$\because$ $a > 0$ and $b > 0$, $\mu_p\left(1-\mu_p\right)-\sigma_p^2 > 0$ $\qquad \therefore\ Var\left[\frac{x}{n}\,|\,n\right] > \sigma_p^2$

E.1.2 Binomial Recovery Model with Seeding Error

This section addresses the situation in which analytical error is modelled by a binomial distribution and in which analytical recovery varies randomly with mean μ_p and variance $\sigma_p{}^2$ and the number of seeded particles varies randomly with mean μ_n and variance $\sigma_n{}^2$. The results of this analysis are discussed in Section 8.1.1 and are summarized by Equations 8.2 and 8.3.

Definitions

n	number of seeded analyte particles
p	analytical recovery of the enumeration method
x	count of analyte particles

Model

$$x \sim BINOMIAL(n, p) \qquad E[x\,|\,n, p] = np \qquad Var[x\,|\,n, p] = np(1-p)$$

Variance Decomposition to Address Non-constant Analytical Recovery

$$E[x\,|\,n] = E\left[E[x\,|\,n, p]\right] = nE[p] = n\mu_p$$

$$Var[x\,|\,n] = E\left[Var[x\,|\,n, p]\right] + Var\left[E[x\,|\,n, p]\right] = nE[p] - nE\left[p^2\right] + n^2 Var[p] = n\mu_p - n\left(\mu_p^2+\sigma_p^2\right) + n^2\sigma_p^2$$

Variance Decomposition to Address Seeding Error

$$E[x] = E\left[E[x\,|\,n]\right] = \mu_p E[n] = \mu_n\mu_p$$

$$\begin{aligned}
Var[x] &= E\left[Var[x\,|\,n]\right] + Var\left[E[x\,|\,n]\right] \\
&= \mu_p E[n] - \left(\mu_p^2+\sigma_p^2\right)E[n] + \sigma_p^2 E\left[n^2\right] + \mu_p^2 Var[n] \\
&= \mu_p\mu_n - \left(\mu_p^2+\sigma_p^2\right)\mu_n + \sigma_p^2\left(\mu_n^2+\sigma_n^2\right) + \mu_p^2\sigma_n^2 \\
&= \mu_p\mu_n - \left(\mu_p^2+\sigma_p^2\right)\mu_n + \sigma_n^2\left(\mu_p^2+\sigma_p^2\right) + \mu_n^2\sigma_p^2
\end{aligned}$$

$E\left[\frac{x}{\mu_n}\right] = \mu_p$ $\therefore$ x/μ_n is an unbiased estimate of mean analytical recovery

$$Var\left[\frac{x}{\mu_n}\right] = \frac{\mu_p}{\mu_n} - \frac{\left(\mu_p^2 + \sigma_p^2\right)}{\mu_n} + \frac{\sigma_n^2\left(\mu_p^2 + \sigma_p^2\right)}{\mu_n^2} + \sigma_p^2$$

E.1.3 Poisson Recovery Model

This section addresses the situation in which seeding and analytical error are collectively modelled by a Poisson distribution and analytical recovery varies randomly with mean μ_p and variance $\sigma_p{}^2$. The results of this analysis are discussed in Section 8.1.2.

Definitions

λ expected number of seeded analyte particles
p analytical recovery of the enumeration method
x count of analyte particles

Model

$x \sim POISSON(\lambda p)$ $E[x \mid p] = \lambda p$ $Var[x \mid p] = \lambda p$

Variance Decomposition to Address Non-constant Analytical Recovery

$$E[x] = E[E[x \mid p]] = \lambda E[p] = \lambda \mu_p$$
$$Var[x] = E[Var[x \mid p]] + Var[E[x \mid p]] = \lambda E[p] + \lambda^2 Var[p] = \lambda \mu_p + \lambda^2 \sigma_p^2$$

$E\left[\frac{x}{\lambda}\right] = \mu_p$ $\therefore$ x/λ is an unbiased estimate of mean analytical recovery

$$Var\left[\frac{x}{\lambda}\right] = \frac{\mu_p}{\lambda} + \sigma_p^2$$

E.1.4 Binomial Recovery Model with Partial Sample Analysis

This section addresses the situation in which analytical error is modelled by a binomial distribution that includes the effect of partial sample analysis and in which analytical recovery varies randomly with mean μ_p and variance $\sigma_p{}^2$. The results of this analysis are discussed in Section 8.1.3 and are summarized by Equation 8.4.

Definitions

n number of seeded analyte particles
θ fraction of sample enumerated
p analytical recovery of the enumeration method
x count of analyte particles

Model

$x \sim BINOMIAL(n, \theta, p)$ $\qquad E[x \mid p] = n\theta p$ $\qquad Var[x \mid p] = n\theta p(1-\theta p)$

Variance Decomposition to Address Non-constant Analytical Recovery

$$\begin{aligned} E[x] &= E[E[x \mid p]] = n\theta E[p] = n\theta\mu_p \\ Var[x] &= E[Var[x \mid p]] + Var[E[x \mid p]] \\ &= n\theta E[p] - n\theta^2 E[p^2] + n^2\theta^2 Var[p] \\ &= n\theta\mu_p - n\theta^2(\mu_p^2 + \sigma_p^2) + n^2\theta^2\sigma_p^2 \end{aligned}$$

$E\left[\frac{x}{n\theta}\right] = \mu_p$ $\quad \therefore$ $x/(n\theta)$ is an unbiased estimate of mean analytical recovery

$$Var\left[\frac{x}{n\theta}\right] = \frac{\mu_p}{n\theta} - \frac{(\mu_p^2 + \sigma_p^2)}{n} + \sigma_p^2$$

E.1.5 Poisson Recovery Model with Partial Sample Analysis

This section addresses the situation in which seeding and analytical error are collectively modelled by a Poisson distribution that includes the effect of partial sample analysis and in which analytical recovery varies randomly with mean μ_p and variance σ_p^2. The results of this analysis are discussed in Section 8.1.3 and are summarized by Equation 8.5.

Definitions

λ	expected number of seeded analyte particles
θ	fraction of sample enumerated
p	analytical recovery of the enumeration method
x	count of analyte particles

Model

$x \sim POISSON(\lambda\theta p)$ $\qquad E[x \mid p] = \lambda\theta p$ $\qquad Var[x \mid p] = \lambda\theta p$

Variance Decomposition to Address Non-constant Analytical Recovery

$$E[x] = E[E[x \mid p]] = \lambda\theta E[p] = \lambda\theta\mu_p$$

$$Var[x] = E[Var[x \mid p]] + Var[E[x \mid p]] = \lambda\theta E[p] + \lambda^2\theta^2 Var[p] = \lambda\theta\mu_p + \lambda^2\theta^2\sigma_p^2$$

$E\left[\frac{x}{\lambda\theta}\right] = \mu_p$ $\quad \therefore$ $x/(\lambda\theta)$ is an unbiased estimate of mean analytical recovery

$$Var\left[\frac{x}{\lambda\theta}\right] = \frac{\mu_p}{\lambda\theta} + \sigma_p^2$$

E.2 VARIANCE DECOMPOSITION OF PROBABILISTIC CONCENTRATION MODELS

This section provides derivations in support of the variance decomposition analyses in Section 8.2. Three different models and their partial sample analysis counterparts are considered.

E.2.1 Constant Concentration

This section addresses the situation in which random sampling and analytical error are collectively modelled by a Poisson distribution and analytical recovery varies randomly with mean μ_p and variance σ_p^2. The results of this analysis are discussed in Section 8.2.1 (and Schmidt *et al.*, 2008) and are summarized by Equation 8.14.

Definitions

c	concentration of analyte particles in the source
V	sample volume
p	analytical recovery of the enumeration method
x	count of analyte particles

Model

$x \sim POISSON(cVp) \qquad E[x \mid p] = cVp \qquad Var[x \mid p] = cVp$

Variance Decomposition to Address Non-constant Analytical Recovery

$$E[x] = E[E[x \mid p]] = cV \cdot E[p] = cV\mu_p$$

$$Var[x] = E[Var[x \mid p]] + Var[E[x \mid p]] = cV \cdot E[p] + c^2V^2 \cdot Var[p] = cV\mu_p + c^2V^2\sigma_p^2$$

$E\left[\frac{x}{V\mu_p}\right] = c \qquad \therefore\ x/(V\mu_p)$ is an unbiased estimate of concentration

$$Var\left[\frac{x}{V\mu_p}\right] = \frac{c}{V\mu_p} + \frac{c^2\sigma_p^2}{\mu_p^2}$$

E.2.2 Constant Concentration with Partial Sample Analysis

This section addresses the situation in which random sampling and analytical error are collectively modelled by a Poisson distribution that includes the effect of partial sample analysis and in which analytical recovery varies randomly with mean μ_p and variance σ_p^2. The results of this analysis are discussed in Section 8.2.1.

Definitions

c concentration of analyte particles in the source
V sample volume
θ fraction of sample enumerated
p analytical recovery of the enumeration method
x count of analyte particles

Model

$$x \sim POISSON(cV\theta p) \qquad E[x \mid p] = cV\theta p \qquad Var[x \mid p] = cV\theta p$$

Variance Decomposition to Address Non-constant Analytical Recovery

$$E[x] = E[E[x \mid p]] = cV\theta \cdot E[p] = cV\theta\mu_p$$

$$Var[x] = E[Var[x \mid p]] + Var[E[x \mid p]] = cV\theta \cdot E[p] + c^2V^2\theta^2 \cdot Var[p] = cV\theta\mu_p + c^2V^2\theta^2\sigma_p^2$$

$$E\left[\frac{x}{\theta V\mu_p}\right] = c$$ $\therefore$ $x/(\theta V\mu_p)$ is an unbiased estimate of concentration

$$Var\left[\frac{x}{\theta V\mu_p}\right] = \frac{c}{\theta V\mu_p} + \frac{c^2\sigma_p^2}{\mu_p^2}$$

E.2.3 Constant Concentration with Non-random Dispersion

This section addresses the situation in which analyte particles in the source are non-randomly dispersed. A hypothetical model is used in which $E(x) = cVp$ and $Var(x) = \delta cVp$. Analytical recovery is assumed to vary randomly with mean μ_p and variance σ_p^2. The results of this analysis are discussed in Section 8.2.2.

Definitions

c concentration of analyte particles in the source
V sample volume
p analytical recovery of the enumeration method
δ dispersion coefficient
x count of analyte particles

Model

$$E[x \mid p] = cVp \qquad Var[x \mid p] = \delta cVp$$

Variance Decomposition to Address Non-constant Analytical Recovery

$$E[x] = E[E[x \mid p]] = cV \cdot E[p] = cV\mu_p$$

$$Var[x] = E[Var[x \mid p]] + Var[E[x \mid p]] = \delta cV \cdot E[p] + c^2V^2 \cdot Var[p] = \delta cV\mu_p + c^2V^2\sigma_p^2$$

$$E\left[\frac{x}{V\mu_p}\right] = c \qquad \therefore\ x/(V\mu_p) \text{ is an unbiased estimate of concentration}$$

$$Var\left[\frac{x}{V\mu_p}\right] = \frac{\delta c}{V\mu_p} + \frac{c^2\sigma_p^2}{\mu_p^2}$$

E.2.4 Variable Concentration

This section addresses the situation in which random sampling and analytical error are collectively modelled by a Poisson distribution and in which analytical recovery varies randomly with mean μ_p and variance $\sigma_p{}^2$ and the concentration varies randomly among samples with mean μ_c and variance $\sigma_c{}^2$ The results of this analysis are discussed in Section 8.2.3 and are summarized by Equation 8.19.

Definitions

c	concentration of analyte particles in the source
V	sample volume
p	analytical recovery of the enumeration method
x	count of analyte particles

Model

$$x \sim POISSON(cVp) \qquad E[x \mid c, p] = cVp \qquad Var[x \mid c, p] = cVp$$

Variance Decomposition to Address Non-constant Analytical Recovery

$$E[x \mid c] = E[E[x \mid c, p]] = cV \cdot E[p] = cV\mu_p$$

$$Var[x \mid c] = E[Var[x \mid c, p]] + Var[E[x \mid c, p]] = cV \cdot E[p] + c^2V^2 \cdot Var[p] = cV\mu_p + c^2V^2\sigma_p^2$$

Variance Decomposition to Address Variability in Concentration

$$E[x] = E[E[x \mid c]] = V\mu_p \cdot E[c] = \mu_p V\mu_c$$

$$\begin{aligned} Var[x] &= E[Var[x \mid c]] + Var[E[x \mid c]] \\ &= \mu_p V \cdot E[c] + V^2\sigma_p^2 \cdot Var\left[c^2\right] + \mu_p^2 V^2 \cdot Var[c] \\ &= \mu_p V\mu_c + V^2\sigma_p^2\left(\mu_c^2 + \sigma_c^2\right) + \mu_p^2 V^2\sigma_c^2 \end{aligned}$$

$$E\left[\frac{x}{V\mu_p}\right] = \mu_c \qquad \therefore\ x/(V\mu_p) \text{ is an unbiased estimate of mean concentration}$$

$$Var\left[\frac{x}{V\mu_p}\right] = \frac{\mu_c}{V\mu_p} + \frac{\sigma_p^2\left(\mu_c^2 + \sigma_c^2\right)}{\mu_p^2} + \sigma_c^2$$

E.3 COMPARISON OF ERRORS ASSOCIATED WITH ANALYTICAL RECOVERY AND INFECTIVITY

This section provides derivations in support of the variance decomposition analyses in Section 6.5.2 that address the assumption that imperfect analytical recovery of pathogens and incomplete infectivity of observed pathogens are counter-balancing errors. The results of this analysis are summarized by Equations 6.13 and 6.14.

Definitions

c	concentration of pathogens in the source
V	sample volume
p	analytical recovery of the enumeration method
x	count of pathogens
i	fraction of the pathogens that are infectious
c_I	concentration of infectious pathogens in the source

Model

$$x \sim POISSON(cVp) \qquad E[x \mid c, p] = cVp \qquad Var[x \mid c, p] = cVp$$

$$c_I = i \times c \qquad E[c_I \mid c, i] = ic \qquad Var[c_I \mid c, i] = 0$$

Variance Decomposition for the Concentration Estimate *x/V*

$$E[x \mid c] = E[E[x \mid c, p]] = cV \cdot E[p] = cV\mu_p$$

$$Var[x \mid c] = E[Var[x \mid c, p]] + Var[E[x \mid c, p]] = cV \cdot E[p] + c^2V^2 \cdot Var[p] = cV\mu_p + c^2V^2\sigma_p^2$$

$$E[x] = E[E[x \mid c]] = V\mu_p \cdot E[c] = V\mu_p\mu_c$$

$$\begin{aligned} Var[x] &= E[Var[x \mid c]] + Var[E[x \mid c]] \\ &= V\mu_p \cdot E[c] + V^2\sigma_p^2 \cdot E[c^2] + V^2\mu_p^2 \cdot Var[c] \\ &= V\mu_p\mu_c + V^2\sigma_p^2\left(\mu_c^2 + \sigma_c^2\right) + V^2\mu_p^2\sigma_c^2 \end{aligned}$$

$$E\left[\frac{x}{V}\right] = \mu_p\mu_c$$

$$Var\left[\frac{x}{V}\right] = \frac{\mu_p\mu_c}{V} + \sigma_p^2\left(\mu_c^2 + \sigma_c^2\right) + \mu_p^2\sigma_c^2$$

Variance of c_I

$$E[c_I \mid c] = E[E[c_I \mid c, i]] = c \cdot E[i] = c\mu_i$$
$$Var[c_I \mid c] = E[Var[c_I \mid c, i]] + Var[E[c_I \mid c, i]] = 0 + c^2 \cdot Var[i] = c^2 \cdot \sigma_i^2$$

$$E[c_I] = E[E[c_I \mid c]] = \mu_i \cdot E[c] = \mu_i \mu_c$$
$$Var[c_I] = E[Var[c_I \mid c]] + Var[E[c_I]] = \sigma_i^2 \cdot E[c^2] + \mu_i^2 \cdot Var[c] = \sigma_i^2 \left(\mu_c^2 + \sigma_c^2\right) + \mu_i^2 \sigma_c^2$$

Notes

Development of a Gibbs sampling algorithm is complicated by the conditional posterior distributions of *a* and *b*. An algorithm to sample from these distributions (approximately) is addressed in Section 4.6.3. This model can be used when replicate recovery data have a mixture of precisely known and Poisson-distributed seed doses simply by fixing the value of n_i for samples with known seed doses. It is possible to combine the Poisson and binomial distributions into x_i~Poisson($\lambda_i p_i$), but the resulting conditional posterior distribution for p_i is complicated.

GLOSSARY

Analytical Error – the difference between the number of target microorganisms or particles observed in a sample and the number that were actually present

Analytical Recovery – "the capacity of the analyst to successfully count each microorganism or particle of interest in a sample using a specific enumeration method" (Schmidt *et al.*, 2010a)

Counting Errors – the difference between the number of enumerable target microorganisms or particles in a processed sample and the count that is obtained by the analyst

Dispersion – the spatial distribution of target microorganisms or particles in a locally homogeneous source (*e.g.* random or clumped)

Measurement Error – the difference between the actual value of a parameter and the value that is estimated from the data (*e.g.* the difference between the actual analytical recovery value and the estimated recovery, or the difference between the actual concentration in the source from which a sample is taken and the concentration that is estimated by enumeration)

Non-constant Analytical Recovery – the variation in analytical recovery among samples

Partial Sample Analysis – the difference between the number of target microorganisms or particles present in a sample and the number that are subjected to the enumeration procedure (due to sub-sampling)

Random Sampling Error – the randomness of the number of target microorganisms or particles contained in a sample from a source with a locally homogeneous concentration because microorganisms and particles are discrete

Replication – the collection of repeated samples that are believed to be representative of the same source concentration or repeated recovery estimates that are believed to be representative of the same non-constant analytical recovery distribution

Sensitivity – the probability that a detection or enumeration method will detect target microorganisms or particles when they are present

Temporal Concentration Variability – the variability in concentration among temporally distributed sampling events at a specific location

References

Ahrens, J.H., Dieter, U., 1974. Computer methods for sampling from gamma, beta, Poisson and binomial distributions. Computing 12(3), 223-246.

Allen, M.J., Clancy, J.L., Rice, E.W. 2000. The plain hard truth about pathogen monitoring. J. AWWA 92(9), 64-76.

Ang, A.H-S., Tang, W.H., 1975. Probability Concepts in Engineering Planning and Design, Volume 1. John Wiley and Sons, Inc., New York.

Assavasilavasukul, P, Lau, B.L.T., Harrington, G.W., Hoffman, R.M., Borchardt, M.A., 2008. Effect of pathogen concentrations on removal of *Cryptosporidium* and *Giardia* by conventional drinking water treatment. Water Res. 42(68), 2678-2690.

Atherholt, T.B., Korn, L.R., 1999. ICR protocol: Alternative treatment of parasite sample data. J. AWWA 91(3), 95-102.

Barbeau, B., Payment, P., Coallier, J., Clément, B., Prévost, M., 2000. Evaluating the risk of infection from the presence of *Giardia* and *Cryptosporidium* in drinking water. Quant. Microbiol. 2(1), 37-54.

Benke and Hamilton, 2008

Box, G.E.P., 1953. Non-normality and tests on variances. Biometrika 40(3-4), 318-335.

Brown, R.A., Cornwell, D.A., 2007. Using spore removal to monitor plant performance for *Cryptosporidium* removal. J. AWWA 99(3), 95-109.

Bukhari, Z., McCuin, R.M., Fricker, C.R., Clancy, J.L., 1998. Immunomagnetic separation of *Cryptosporidium parvum* from source water samples of various turbidities. Appl. Environ. Microbiol. 64(11), 4495-4499.

Chae, G-T., Stimson, J., Emelko, M.B., Blowes, D.W., Ptacek, C.J., Mesquita, M.M., 2008. Statistical assessment of the accuracy and precision of bacteria- and virus-sized microsphere enumerations by epifluorescence microscopy. Water Res. 42(6-7), 1431-1440.

Christian, R.R., Pipes, W.O. 1983. Frequency distribution of coliforms in water distribution systems. Appl. Environ. Microbiol. 45(2), 603-609.

Clancy, J.L., Gollnitz, W.D., Tabib, Z., 1994. Commercial labs: How accurate are they? J. AWWA 86(5), 89-97.

Clancy, J.L., Connell, K., McCuin, R.M., 2003. Implementing PBMS improvements to USEPA's *Cryptosporidium* and *Giardia* methods. J. AWWA 95(9), 80-93.

Cook, N., Paton, C.A., Wikinson, N., Nichols, R.A.B., Barker, K., Smith, H.V., 2006. Towards standard methods for the detection of *Cryptosporidium parvum* on lettuce and raspberries. Part 1: Development and optimization of methods. Int. J. Food Microbiol. 109(3), 215-221.

Crainiceanu, C.M., Stedinger, J.R., Rupert, D., Behr, C.T., 2003. Modeling the U.S. national distribution of waterborne pathogen concentrations with application to *Cryptosporidium parvum*. Water Resour. Res. 39(9), 1235-1249.

Craun, G.F., Hubbs, S.A., Frost, F., Calderon, R.L., Via, S.H., 1998. Waterborne outbreaks of cryptosporidiosis. J. AWWA 90(9), 81-91.

Crockett, C.S., Haas, C.N., 1995. Protozoan monitoring: From the ICR to the ESWTR. J. AWWA 87(8), 50-59.

Cummins, E., Kennedy, R., Cormican, M., 2010. Quantitative risk assessment of *Cryptosporidium* in tap water in Ireland. Sci. Total Environ. 408(4), 740-753.

Davies, C.M., Altavilla, N., Krogh, M., Ferguson, C.M., Deere, D.A., Ashbolt, N.J., 2005. Environmental inactivation of *Cryptosporidium* oocysts in catchment soils. J. Appl. Microbiol. 98(2), 308-317.

DiGiorgio, C.L., Gonzales, D.A., Huitt, C.C., 2002. *Cryptosporidium* and *Giardia* recoveries in natural waters by using Environmental Protection Agency Method 1623. Appl. Environ. Microbiol. 68(12), 5952-5955.

Downey, A.S., Graczyk, T.K., 2007. Maximizing recovery and detection of *Cryptosporidium parvum* oocysts from spiked Eastern oyster (*Crassostrea virginica*) tissue samples. Appl. Environ. Microbiol. 73(21), 6910-6915.

Dugan, N.R., Fox, K.R., Owens, J.H., Miltner, R.J., 2001. Controlling *Cryptosporidium* oocysts using conventional treatment. J. AWWA 93(12), 64-76.

DuPont, H.L., Chappell, C.L., Sterling, C.R., Okhuysen, P.C., Rose, J.B., Jakubowski, W., 1995. The infectivity of *Cryptosporidium parvum* in healthy volunteers. New Engl. J. Med. 332(13), 855-859.

Edmonds, J.M., Collett, P.J., Valdes, E.R., Skowronski, E.W., Pellar, G.J., Emanuel, P.A., 2009. Surface sampling of spores in dry-deposition aerosols. Appl. Environ. Microbiol. 75(1), 39-44.

Eisenhart, C, Wilson, P.W., 1943. Statistical methods and control in bacteriology. Bacteriol. Rev. 7(2), 57-137.

El-Shaarawi, A.H., Esterby, S.R., Dutka, B.J., 1981. Bacterial density in water determined by Poisson or negative binomial distributions. Appl. Environ. Microbiol. 41(1), 107-116.

Emelko, M.B., 2001. Removal of *Cryptosporidium parvum* by granular media filtration. Ph.D. Dissertation, University of Waterloo. Waterloo, Ontario, Canada.

Emelko, M.B., Huck, P.M., 2004. Microspheres as surrogates for *Cryptosporidium* filtration. J. AWWA 96(3), 94-105.

Emelko, M.B., Schmidt, P.J., Roberson, J.A., 2008. Quantification of uncertainty in microbial data – reporting and regulatory implications. J. AWWA 100(3), 94-104.

Emelko, M.B., Schmidt, P.J., Reilly, P.M., 2010a. Particle and microorganism enumeration data: Enabling quantitative rigor and judicious interpretation. Environ. Sci. Technol. 44(5), 1720-1727.

Emelko, M.B., Schmidt, P.J., Reilly, P.M., 2010b. Particle and microorganism enumeration data: Enabling quantitative rigor and judicious interpretation. Supporting Information (available online). Environ. Sci. Technol. 44(5), 1720-1727.

Evans, M., Hastings, N., Peacock, B., 2000. Statistical distributions; 3rd ed. John Wiley and Sons, Inc., New York.

Falk, C.C., Karanis, P., Schoenen, D., Seitz, H.M., 1998. Bench scale experiments for the evaluation of a membrane filtration method for the recovery efficiency of *Giardia* and *Cryptosporidium* in water. Water Res. 32(3), 565-568.

Feng, Y.Y., Ong, S.L., Hu, J.Y., Song, L.F., Tan, X.L., Ng, W.J., 2003. Effect of particles on the recovery of *Cryptosporidium* oocysts from source water samples of various turbidities. Appl. Environ. Microbiol. 69(4), 1898-1903.

Ferguson, C., Kaucner, C., Krogh, M., Deere, D., Warnecke, M., 2004. Comparison of methods for the concentration of *Cryptosporidium* oocysts and *Giardia* cysts from raw waters. Can. J. Microbiol. 50(9), 675-682.

Fisher, R.A., Thornton, H.G., MacKenzie, W.A., 1922. The accuracy of the plating method of estimating the density of bacterial populations. With particular reference to the use of Thornton's agar medium with soil samples. Ann. Appl. Biol. 9(3), 325-359.

Fisher, R.A., 1925. Theory of statistical estimation. Proc. Cambridge Philos. Soc. 22, 700-725.

Fisher, R.A., 1941. The negative binomial distribution. Ann. Eugenic. 11(S), 182-187.

Franceschina, C.C., Schmidt, P.J., Emelko, M.B., Graeff-Teixeira, C., In Progress. Improving method validation by understanding errors in the preparation of standards: Introducing low numbers of *Schistosoma mansoni* eggs into feces.

Furumoto, W.A., Mickey, R., 1967. Mathematical model for the infectivity-dilution curve of tobacco mosaic virus: theoretical considerations. Virology 32(2), 216-223.

Gale P., 1996. Developments in microbiological risk assessment models for drinking water – a short review. J. Appl. Bacteriol. 81(4), 403-410.

Gale, P., van Dijk, P.A.H., Stanfield, G., 1997. Drinking water treatment increases microorganism clustering: The implications for microbiological risk assessment. J. Water Supply Res. Technol. – AQUA 46(3), 117-126.

Gale P., 1998. Simulating *Cryptosporidium* exposures in drinking water during an outbreak. Water Sci. Technol. 38(12), 7-13.

Gelman A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis, 2nd ed. Chapman & Hall/CRC, London.

Graeff-Teixeira, C., 2008. Personal Communication.

Haas, C.N., 1983. Estimation of risk due to low doses of microorganisms: A comparison of alternative methodologies. Am. J. Epidemiol, 118(4), 573-582.

Haas, C.N., 1993. Microbial sampling: Is it better to sample many times or use large samples? Water Sci. Technol. 27(3-4), 19-25.

Haas, C.N., 1996. How to average microbial densities to characterize risk. Water Res. 30(4), 1036-1038.

Haas, C.N., 2002. Progress and data gaps in quantitative microbial risk assessment. Water Sci. Technol. 46(11-12), 277-284.

Haas, C.N., Heller, B., 1986. Statistics of enumerating total coliforms in water samples by membrane filter procedures. Water Res. 20(4), 525-530.

Haas, C.N., Heller, B., 1988. Test of the validity of the Poisson assumption for analysis of Most-Probable-Number results. Appl. Environ. Microbiol. 54(12), 2996-3002.

Haas, C.N., Rose, J.B., 1996. Distribution of *Cryptosporidium* oocysts in a water supply. Water Res. 30(10), 2251-2254.

Haas, C.N., Rose, J.B., Gerba, C., Regli, S., 1993. Risk assessment of virus in drinking water. Risk Anal. 13(5), 545-552.

Haas, C.N., Crockett, C.S., Rose, J.B., Gerba, C.P., Fazil, A.M., 1996. Assessing the risk posed by oocysts in drinking water. J. AWWA 88(9), 131-136.

Haas, C.N., Rose, J.B., Gerba, C.P., 1999. Quantitative Microbial Risk Assessment. John Wiley and Sons, New York.

Hansen, J.S., Ongerth, J.E., 1991. Effects of time and watershed characteristics on the concentration of *Cryptosporidium* oocysts in river water. Appl. Environ. Microbiol. 57(10), 2790-2795.

Havelaar, A.H., De Hollander, A.E.M., Teunis, P.F.M., Evers, E.G., Van Kranen, H.J., Versteegh, J.F.M., Van Koten, J.E.M., Slob, W., 2000 Balancing the risks and benefits of drinking water disinfection: disability adjusted life-years on the scale. Environ. Health Persp. 108(4), 315–321.

Huck, P.M., Emelko, M.B., Coffey, B.M., Maurizio, D.D., O'Melia, C.R., 2001 Filter Operation Effects on Pathogen Passage. AWWA Research Foundation and American Water Works Association.

Hobert, J.P., Casella, G., 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. J. Am. Statist. Assn. 91(436), 1461-1473.

Iacovski, R.B., Barardi, C.R.M., Simões, C.M.O., 2004. Detection and enumeration of *Cryptosporidium* sp. oocysts in sewage sludge samples from the city of Florianópolis (Brazil) by using immunomagnetic separation combined with indirect immunofluorescence assay. Waste Manage. Res. 22(3), 171-176.

Jaidi, K., Barbeau, B., Carrière, A., Desjardins, R., Prévost, M., 2009. Including operational data in QMRA model: development and impact of model inputs. J. Water Health 7(1), 77-95.

Jeffreys, H., 1961. Theory of Probability, 3rd ed. Oxford University Press, London.

Katz, N., Chaves, A., Pellegrino, J., 1972. A simple device for quantitative stool thick-smear technique in schistosomiasis mansoni. Rev. Inst. Med. trop. S. Paulo. 14(6), 397-400.

Kinzelman, J., Ng, C., Jackson, E., Gradus, S., Bagley, R., 2003. Enterococci as indicators of Lake Michigan recreational water quality: Comparison of two methodologies and their impacts on public health regulatory events. Appl. Environ. Microbiol. 69(1), 92-96.

Korich, D.G., Mead, J.R., Madore, M.S., Sinclair, N.A., Sterling, C.R. 1990. Effects of ozone, chlorine dioxide, chlorine, and monochloramine on *Cryptosporidium parvum* oocyst viability. Appl. Environ. Microbiol. 56(5), 1423-1428.

Kuczynska, E., Shelton, D.R., 1999. Method for detection and enumeration of *Cryptosporidium parvum* oocysts in feces, manures, and soils. Appl. Environ. Microbiol. 65(7), 2820-2826.

LeChevallier, M.W., Norton, W.D., Lee, R.G., 1991a. Occurrence of *Giardia* and *Cryptosporidium* spp. in surface water supplies. Appl. Environ. Microbiol., 57(9), 2610-2616.

LeChevallier, M.W., Norton, W.D., Lee, R.G., 1991b. *Giardia* and *Cryptosporidium* spp. in filtered drinking water supplies. Appl. Environ. Microbiol., 57(9), 2617-2621.

LeChevallier, M.W., Norton, W.D., Siegel, J.E., Abbaszadegan, M., 1995. Evaluation of the immunofluorescence procedure for detection of *Giardia* cysts and *Cryptosporidium* oocysts in water. Appl. Environ. Microbiol. 61(2), 690-697.

Margolin, B.H., Kaplan, N., Zeiger, E., 1981. Statistical analysis of the Ames *Salmonella*/microsome test. Proc. Natl. Acad. Sci. USA 78(6), 3779-3783.

Masago, Y., Katayama, H., Hashimoto, A., Hirata, T., Ohgaki, S., 2002. Assessment of risk of infection due to *Cryptosporidium parvum* in drinking water. Water Sci. Technol. 66(11-12), 319-324.

Massanet-Nicolau, J., 2003. New method using sedimentation and immunomagnetic separation for isolation and enumeration of *Cryptosporidium parvum* oocysts and *Giardia lamblia* cysts. Appl. Environ. Microbiol. 69(11), 6758-6761.

McCuin, R.M., Clancy, J.L., 2003. Modifications to USEPA Methods 1622 and 1623 for detection of *Cryptosporidium* oocysts and *Giardia* cysts in water. Appl. Environ. Microbiol. 69(1), 267-274.

McCuin, R.M., Bukhari, Z., Sobrinho, J., Clancy, J.L., 2001. Recovery of *Cryptosporidium* oocysts and *Giardia* cysts from source water concentrates using immunomagnetic separation. J. Microbiol. Methods 45(2), 69-76.

Medema, G.J., Teunis, P.F.M., Gornik, V., Havelaar, A.H., Exner, M., 1995. Estimation of the *Cryptosporidium* infection risk via drinking water. In "Protozoa Parasites and Water" edited by Betts, W.B., Casemore, D., Fricker, C., Smith, H., Watkins, J. The Royal Society of Chemistry, pp. 53-56.

Medema, G.J., Hoogenboezem, W., van der Veer, A.J., Ketelaars, H.A.M., Hijnen, W.A.M., Nobel, P.J., 2003. Quantitative risk assessment of *Cryptosporidium* in surface water treatment. Water Sci. Technol. 47(3), 241-247.

Mons, C., Dumètre, A., Gosselin, S., Galliot, C., Moulin, L., 2009. Monitoring of *Cryptosporidium* and *Giardia* river contamination in Paris area. Water Res. 43(1), 211-217.

Nahrstedt, A., Gimbel, R., 1996. A statistical method for determining the reliability of the analytical results in the detection of *Cryptosporidium* and *Giardia* in water. J. Water Supply Res. Technol. – Aqua 45(3), 101-111.

Nieminski, E., Schaefer, F.W., Ongerth, J.E., 1995. Comparison of two Methods for Detection of *Giardia* Cysts and *Cryptosporidium* Oocysts in Water. Appl. Environ. Microbiol., 61(5), 1714-1719.

Parkhurst, D.F., Stern, D.A., 1998. Determining average concentrations of *Cryptosporidium* and other pathogens in water. Environ. Sci. Technol. 32(21), 3424-3429.

Parks, S.L.I., VanBriesen, J.M., 2009. Evaluating temporal variability in bacterial indicator samples for an urban watershed. J. Environ. Eng. 135(12), 1294-1303.

Peeters, J.E., Mazas, E.A., Masschelein, W.J., De Maturana, I.V.M., Debacker, E., 1989. Effect of disinfection of drinking water with ozone or chlorine dioxide on survival of *Cryptosporidium parvum* oocysts. Appl. Environ. Microbiol. 55(6), 1519-1522.

Pereira, M.D.G., Atwill, E.R., Jones, T., 1999. Comparison of sensitivity of immunofluorescent microscopy to that of a combination of immunofluorescence microscopy and immunomagnetic separation for detection of *Cryptosporidium parvum* oocysts in adult bovine feces. Appl. Environ. Microbiol. 65(7), 3236-3239.

Petterson, S.R., Signor, R.S., Ashbolt, N.J., 2007. Incorporating method recovery uncertainties in stochastic estimates of raw water protozoan concentrations for QMRA. J. Water Health 5(S1), 51-65.

Pipes, W.O., Ward, P., Ahn, S.H., 1977. Frequency distribution for coliform bacteria in water. J. AWWA 69(12), 664-668.

Pouillot, R., Beaudeau, P., Denis, J., Derouin, F., 2004. A quantitative risk assessment of waterborne cryptosporidiosis in France using second-order Monte Carlo simulation. Risk Anal. 24(1), 1-17.

Prévost, M., Barbeau, B., 2010. Personal Communication.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C: The art of scientific computing, 2nd ed. Cambridge University Press, Cambridge.

Regli, S., Rose, J.B., Haas, C.N., Gerba, C.P., 1991. Modeling the risk from *Giardia* and viruses in drinking water. J. AWWA 83(9), 76-84.

Rendtorff, R.C., 1954. The experimental transmission of human intestinal protozoan parasites. II. *Giardia lamblia* cysts given in capsules. Am. J. Hyg. 59(2), 209-220.

Reynolds, D.T., Slade, R.B., Sykes, N.J., Jonas, A., Fricker, C.R., 1999. Detection of *Cryptosporidium* oocysts in water: techniques for generating precise recovery data. J. Appl. Microbiol. 1999, 87(6), 804-813.

Rice, E.W., Fox, K.R., Miltner, R.J., Lytle, D.A., Johnson, C.H., 1996. Evaluating plant performance with endospores. J. AWWA 88(9), 122-130.

Roberson, J.A., Schmidt, P.J., Emelko, M.B., 2010. Using statistical models to transform microbial monitoring data into information for decision-makers. AWWA Int'l Symp. on Waterborne Pathogens, American Water Works Association, Denver.

Robinson, G., Watkins, J., Chalmers, R.M., 2008. Evaluation of a modified semi-automated immunomagnetic separation technique for the detection of *Cryptosporidium* oocysts in human faeces. J. Microbiol. Methods 75(1), 139-141.

Rose, J.B., Haas, C.N., Regli, S., 1991a. Risk assessment and control of waterborne giardiasis. Am. J. Public Health 81(6), 709-713.

Rose, J.B., Gerba, C.P., 1991. Use of risk assessment for development of microbial standards. Water Sci. Technol. 24(2), 29-34.

Rose, J.B., Gerba, C.P., Jakubowski, W., 1991b. Survey of potable water supplies for *Cryptosporidium* and *Giardia*. Environ. Sci. Technol. 25(8), 1393-1400.

Rose, J.B., Lisle, J.T., Haas, C.N., 1995. Risk assessment methods for *Cryptosporidium* and *Giardia* in contaminated water. In "Protozoa Parasites and Water" edited by Betts, W.B., Casemore, D., Fricker, C., Smith, H., Watkins, J. The Royal Society of Chemistry, pp. 238-242.

Rosen, J.S., Sobrinho, J.A.H., LeChevallier, M., 2009. Statistical limitations in the usefulness of total coliform data. J. AWWA 101(3), 68-81.

Sartor, M., Antonena, V., Garvin, F., Webb, M., Bradstock, K.F., 2005. Recovery of viable CD34$^+$ cells from cryopreserved hemopoietec progenitor cell products. Bone Marrow Trans. 36(3), 199-204.

Schmidt, P.J., Emelko, M.B., 2007. Statistical analysis of microbial data obtained by serial dilution – agar plate methods. AWWA Water Quality Technology Conference, American Water Works Association, Denver.

Schmidt, P.J., Emelko, M.B., 2010a. QMRA and decision-making: Are we handling measurement errors associated with pathogen concentration data correctly? Water Res. doi: 10.1016/j.watres.2010.08.042.

Schmidt, P.J., Emelko, M.B., 2010b. QMRA and decision-making: Are we handling measurement errors associated with pathogen concentration data correctly? Supplementary Content (available online). Water Res. doi: 10.1016/j.watres.2010.08.042.

Schmidt, P.J., Emelko, M.B., Thompson, M.E., 2008. Getting better microbial data: Strategies to quantify and reduce uncertainty. AWWA Water Quality Technology Conference, American Water Works Association, Denver.

Schmidt, P.J., Emelko, M.B., Reilly, P.M., 2010a. Quantification of analytical recovery in particle and microorganism enumeration methods. Environ. Sci. Technol. 44(5), 1705-1712.

Schmidt, P.J., Emelko, M.B., Reilly, P.M., 2010b. Quantification of analytical recovery in particle and microorganism enumeration methods. Supporting Information (available online). Environ. Sci. Technol. 44(5), 1705-1712.

Shepherd, K.M., Wyn-Jones, A.P., 1995. Evaluation of different filtration techniques for the concentration of *Cryptosporidium* oocysts from water. Water Sci. Technol. 31(5-6), 425-429.

Signor, R.S., Ashbolt, N.J., 2006. Pathogen monitoring offers questionable protection against drinking-water risks: a QMRA (Quantitative Microbial Risk Analysis) approach to assess management strategies. Water Sci. Technol. 54(3), 261-268.

Smeets, P.W.M.H., van Dijk, J.C., Stanfield, G., Rietveld, L.C., Medema, G.J., 2007. How can the UK statutory *Cryptosporidium* monitoring be used for Quantitative Risk Assessment of *Cryptosporidium* in drinking water? J. Water Health 5(S1), 107-118.

Snedecor, G.W., Cochran, W.G., 1989. Statistical Methods, 8th ed. Iowa State University Press, Ames, Iowa.

'Student', 1907. On the error of counting with a haemacytometer. Biometrika 5(3), 351-360.

Teixeira, C.F., Neuhauss, E., Ben, R., Romanzini, J., Graeff-Teixeira, C., 2007. Detection of *Schistosoma mansoni* eggs in feces through their interaction with paramagnetic beads in a magnetic field. PLoS Negl. Trop. Dis. 1(2), e73.

Teunis, P.F.M., Havelaar, A.H., 1999. *Cryptosporidium* in drinking water: Evaluation of the ILSIIRSI quantitative risk assessment framework. RIVM Report no. 284 550 006, RIVM, Bilthoven, The Netherlands.

Teunis, P.F.M., Havelaar, A.H., 2002. Risk assessment for protozoan parasites. Int. Biodeter. Biodegr. 50(3-4), 185-193.

Teunis, P.F.M., Medema, G.J., Kruidenier, L., Havelaar, A.H., 1997. Assessment of the risk of infection by *Cryptosporidium* or *Giardia* in drinking water from a surface water source. Water Res. 31(6), 1333-1346.

Teunis, P.F.M., Evers, E.G., Slob, W., 1999. Analysis of variable fractions resulting from microbial counts. Quant. Microbiol. 1(1), 63-88.

Teunis, P.F.M., Rutjes, S.A., Westrell, T., de Roda Husman, A.M., 2009. Characterization of drinking water treatment for virus risk assessment. Water Res. 43(2), 395-404.

USEPA, 1996. ICR Microbial Laboratory Manual; EPA 600-R-95-178, U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA, 2004. Water quality standards for coastal and Great Lakes recreation waters; Final rule. Fed. Regist. 69(220), 67217-67243.

USEPA, 2005a. Method 1622: *Cryptosporidium* in Water by Filtration/IMS/FA; EPA 815-R-05-001. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA, 2005b. Method 1623: *Cryptosporidium* and *Giardia* in Water by Filtration/IMS/FA; EPA 815-R-05-002. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA, 2005c. Occurrence and Exposure Assessment for the Final Long Term 2 Enhanced Surface Water Treatment Rule; EPA 815-R-06-002, U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA, 2006. Long term 2 enhanced surface water treatment rule; Final rule. Fed. Regist. 71(3), 654-786.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Springer, New York.

Vesey, G., Slade, J.S., Byrne, M., Shepherd, K., Fricker, C.R., 1993. A new method for the concentration of *Cryptosporidium* oocysts from water. J. Appl. Bacteriol. 75(1), 82-86.

Villegas, L.F., 2010. Variations in oocyst recovery associated with the origin of *Cryptosporidium* isolates. AWWA Int'l Symp. on Waterborne Pathogens, American Water Works Association, Denver.

Warnecke, M., Weir, C., Vesey, G., 2003. Evaluation of an internal positive control for *Cryptosporidium* and *Giardia* testing in water samples. Lett. Appl. Microbiol. 37(3), 244-248.

Young, P.L., Komisar, S.J., 1999. The variability introduced by partial sample analysis to numbers of *Cryptosporidium* oocysts and *Giardia* cysts reported under the information collection rule. Water Res. 33(11), 2660-2668.