Mathematical Modeling of the Separation Process of Chromatography and Estimation of Parameters

By

Xueyi Chen

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chairperson, Jonathan D. Mahnken, PhD

_____

Matthew S. Mayo, PhD

_____

Francisco J. Diaz, PhD

_____

Jo A. Wick, PhD

_____

Mark T. Fisher, PhD

Date Defended: Aug. 24, 2016

The Dissertation Committee for Xueyi Chen

Certifies that this is the approved version of the following dissertation:

Mathematical Modeling of the Separation Process of Chromatography and Estimation of Parameters

_____

Chairperson, Jonathan D. Mahnken.

Date approved: Aug. 24, 2016

Abstract

Chromatography is widely used as a technology for separating mixtures of compounds by partitioning into the mobile and stationary phases. A mathematical model is essential not only for predicting the retention time and the peak shape of the chromatography analyte concentration distribution, but also for understanding the separation mechanism of chromatography and detecting whether the conditions were correct (e.g., whether there was an overload of the sample). A variety of statistical distribution functions such as exponential, Gaussian (normal), exponential modified Gaussian, Weibull, log-normal have been used to approximate the chromatography analyte concentration distributions, and were further applied to the deconvolution of stacked peaks.

The dissertation consists of five chapters. The first chapter presents an overall introduction of the current prevailing mathematical models of chromatography analyte concentration distributions, the generalized chromatography theorem derived from chromatography table and its proof, the relation between on-chromatography analyte concentration distributions and out flow analyte concentration distributions, the asymptotic distribution of on-chromatography analyte concentration distributions and out flow analyte concentration distributions and their applications.

The second chapter presents the mathematical model for the separation process of chromatography. In this chapter the first generalized theorem for modeling almost all types of chromatography was developed, and was found to match the mathematical formulas for well-known discrete distribution functions. These empirical formulas were rigorously proven by mathematical induction based on chromatography principle and chromatography process assumptions. The outflow chromatography analyte concentration distributions are demonstrated

by simulation to be better approximated by the mathematical model that matches the negative binomial distribution function versus using a Gaussian distribution function, which currently is widely used for approximation.

The third chapter establishes the mathematic bridge between on-chromatography and outflow analyte concentration distributions. In following with the previous chapter, which found the on chromatography and outflow analyte concentrations distributions to mathematically match the binomial and negative binomial distributions, respectively, this mathematical bridge can apply to relate these statistical distributions given they mathematical formulas are the same. This theorem is rigorously proved by mathematical induction. This relation is also demonstrated by 3D-plot of on-chromatography and outflow analyte concentration distributions for the first several stages.

The fourth chapter proposed the transformation of data collected by chromatography (i.e., the analyte concentration distributions from chromatography experiments) into data that can be used for estimation to the approximate the underlying parameters that govern a particular chromatography process. Outflow chromatography analyte concentration distribution from original data were used to demonstrate this process and to compare the approach derived in this work using parameter estimated by method of moment (MOM) to the currently approach based on the Gaussian statistical distribution's formula

The fifth chapter is the summary of my dissertation work.

Acknowledgement

Firstly, I would like to express my great appreciation to my academic advisor as well as dissertation supervisor, Dr. Jonathan Mahnken, who guided and supported my desertion research with his profound statistic theory and application knowledge. Without him this dissertation might not be accomplished in current form. I would like to express my great appreciation to department of biostatistics and Dr. Matthew S. Mayo's both financial and spiritual support for me to finish both Master and Doctoral degrees in biostatistics. I would also like to thank my committee members, Dr. Matthew S. Mayo, Dr. J. Francisco Diaz, and Dr. Jo A. Wick, Dr. Mark T. Fisher, Dr. Hung-Wen (Henry) Yeh for their guidance of my disertation. . I would also like to thank Dr. Wick and Shana for their help regarding my academic progresses. Lastly, I would like to thank my family for their support of my academic activities as well as degree accomplishments.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Many mixtures appear to be homogeneous. For example if salt and sugar are grinded together, the resulting powder appears to be uniform; however, it is still a combination of two different components. In many occasions people realize that uniform powder may not only be composed of a single pure component and try different methods to separate them such as the recrystallization method developed in the ancient time to separate the table salt from the sea salt.

In early 1900s, chromatography was developed as a convenient and important method to separate different component that were dispersed uniformly in a mixture. A Russian-Italian botanist Mikhail Semyonovich first discovered that different coloured component in plant (plant pigment) can be separated by using liquid chromatography with calcium carbonate as stationary phase and petroleum ether/ethanol mixture as eluent. [1] Chromatography is applied in many fields such as toxicology, environmental science and criminal science investigations as the rigorous method to confirm the existence of a certain chemical compound. For example, gas chromatography (GC) and high pressure liquid chromatography (HPLC) were used to separate and confirm the blood and urine drug metabolites in these human fluid samples. HPLC was also applied in detecting the concentration level of lead, mercury and arsenic in bodies of waters such as lakes, rivers and reservoirs to protect people from potential heavy metal poisonings.

Different compounds have different affinities to the stationary phase as “temporary binder” when traveling through chromatography with mobile phase known as “eluent” and thus they are to be collected at different times in the solution that flows through the chromatography, called the outflow solution. The time a particular compound spent in the chromatography is

called its “retention time”. By partitioning into the mobile and stationary phases, the chromatography is widely applied as technology for separating mixtures of compounds [2].

1.1 Chromatographic Models.

To date, many theoretical and experimental attempts have been made to understand the separation mechanism of chromatography and optimized its conditions [3]. The retention time and the peak shape of the chromatography analyte concentration distribution are the two important factors that researchers would like to be able to predict based on the previous experimental data. Empirical or semi-empirical peak shape-matching have been applied to approximate the chromatography peaks using a variety of mathematical functions that represent statistical distribution functions such as exponential, Gaussian (normal), exponential modified Gaussian, Weibull, log-normal distributions [4]. However, this type of empirical peak shape-matching can cause confusions since the statistical distribution functions and corresponding parameters to match chromatography separation are anecdotal. Moreover, most researchers attempt to pick different statistical distribution functions as model to match the chromatography shape without considering the mechanism of chromatography separation process.

To solve these problems, a universal mathematical model needs to be developed for the separation process of almost all type of chromatography based on chromatography principle and assumptions of the chromatography process. In this work, the first generalized theorem to model the peaks, or analyte concentration distributions, for all types of chromatography using chromatography tables will be proposed. The binomial distribution was the correct numeric function for mathematically modeling on-chromatography analyte concentration distributions, and the negative binomial distribution was the correct numeric formula for mathematically modeling outflow analyte concentration distributions. These proposed conjectures will be

rigorously proved by mathematical induction. Simulations were conducted to demonstrate that outflow analyte concentration distributions governed by our stated chromatography process assumptions are better approximated using the mathematical formula used for the negative binomial distribution in contrast to the widely utilized Gaussian distribution's mathematical formula.

The dispute among researchers regarding what type of mathematical function is more suitable to describe the on-chromatography and outflow analyte concentration distributions went on for decades without agreement. Very few efforts have been made to the understanding of the difference and relationship between on-chromatography analyte concentration distribution and outflow analyte concentration distribution [5], which is the key to solve this dispute. We have established the generalized theorem of the chromatography model: on-chromatography analyte concentration distributions follow the same mathematical formula as the binomial distribution, and outflow analyte concentration distributions follow the same mathematical formula as the negative binomial distribution.

The binomial and negative binomial distributions have been of both theoretical and application interest for decades [6]. However, outside of their probabilistic context where they are defined by a series of independent Bernoulli trials, the relation between mathematical formulas for the binomial and negative binomial distributions is not clear. In this work the relation between the mathematical formulas binomial and negative binomial distribution will be unambiguously established and rigorously proved using mathematical induction—outside of the context of probability. The chromatography analyte concentration distribution for the first several stages will be plotted in 3 dimensions, which assists the visualization of this relationship.

The Gaussian distribution is the current prevailing numerical model to approximate many types of chromatography by empirically matching observed peaks to match. It was not justified by the chromatography principle and mechanism of the chromatography separation process. In this work, we have proved that as number of theoretical plates approaches infinity, both on-chromatography and outflow analyte concentration distributions, which mathematically match the binomial and negative binomial distributions, respectively, approaches (mathematically) the Gaussian distribution. Thus, using the mathematical formula of the Gaussian distribution to approximate the analyte concentration distribution is appropriate for chromatography with continuous measures/peaks such as gas chromatography (GC), high pressure liquid chromatography (HPLC). (Continuous peaks are defined by the equation $\lim_{t\to c} h(t) = h(c)$ where t represents the time, h represents the analyte concentration distribution height as a function of time and c is constant.)

The chromatography analyte concentration distributions are usually reproducible if research retain same conditions, thus the parameters of chromatography are relatively constant across multiple separation process. The estimation of chromatography parameter is of research interest with many applications such as chromatography peak (analyte concentration distribution) simulation, analyte component selection and multiple peaks deconvolution [7]. To date, chromatography parameters are determined by separate sets of experiments [8], which is costly and time consuming. In this work, we develop a transform method that converts measured chromatography analyte concentration distributions into the form of data commonly collected as statistical sampling data so that statistical methods can be used for parameter estimation. Lastly, this transformation method was applied to convert a set of experimental data collected during a chemical compound separation experiment into the form of statistical sampling data, and the

chromatography parameters were then estimated. Simulated outflow peaks based on estimated chromatography parameters better resembled the outflow peak plot from the observed experimental data as compared to current prevailing Gaussian peak matching method.

1.2 Definition of Chromatographic Terms.

Mobile phase is the phase that moves in a direction toward outlet

Stationary phase is the materials fixed in position during the separation process for the chromatography

Note: Mobile phase and stationary phase are the two phases in chromatography

Analytes are the substances to be separated through the chromatographic process

Note: The analytes are partitioned in both mobile phase and stationary phase governed by specific partition coefficients.

Analytes concentration distribution is the curve representing analyte concentration versus the stage.

Note: The general term "distribution" refers to a probabilistic or statistical distribution, whereas the term "analytes concentration distribution" refers to the chromatography analyte concentration distribution curve, which is commonly described as distribution of chromatography peak in field of chromatography separation

Eluent is the solvent that dissolve the analytes (solute) and it flows in mobile phase carrying analytes down the stream to the outlet of the chromatography.

Partition coefficient ρ is a constant ratio of the concentration of an analyte in mobile phase to the concentration of this analyte in stationary phase at equilibrium.

Proportion constant λ is proportion of the analyte in mobile phase at equilibrium, and

$$\lambda = \frac{\rho}{1+\rho}$$

On chromatography denotes the analyte that is still on the chromatography column.

Outflow denotes the state that analyte flow out of chromatography system.

Retention time is the length of time an analyte is retained on a chromatography column.

Dead volume is the total volume of the liquid phase in the chromatographic column. It is a parameter which is independent of the types of analytes and mobile phase.

Chapter 2: Mathematical Model for the Separation Process of Chromatography

2.1. Introduction

Chromatography is widely used as a technology for separating mixtures of compounds by partitioning into the mobile and stationary phases.[9] A mathematical model is essential not only for predicting the retention time and the peak shape of the chromatography analyte concentration distribution, but also for understanding the separation mechanism of chromatography and detecting whether the conditions were correct (e.g., whether there was an overload (excessive amount of analyte was added to the chromatography system) of the sample.[10] A variety of statistical distribution functions such as exponential, Gaussian (normal), exponential modified Gaussian, Weibull, log-normal have been used to approximate the chromatography analyte concentration distributions, and were further applied to the deconvolution of stacked peaks[10][11]. However, the empirical or semi-empirical "peak shape-matching" (i.e., matching the chromatography analyte concentration distributions) to the mathematical formulas from known statistical distribution functions have been anecdotal, and as a result the empirically chosen statistical distribution functions and corresponding parameters applied within the context of the chromatography separation can cause confusion when utilizing the statistical terminology. It also cannot be stated with sufficient confidence that chromatographic models based on matching peak shapes or numerical simulations generalize to all types of chromatography. Furthermore, although it had been reported that the outflow chromatography peaks often followed skewed distributions [12], in many cases researchers still apply use mathematical formulas for distributions such as Gaussian, and other symmetric

distributions to model outflow chromatographic peaks analyte concentration distributions and conduct deconvolutions based on these symmetric distributions formulas. To date very few studies had been conducted on understanding the intrinsic mechanism of chromatography separation and distinguish the statistical distribution function formulas used s for on-chromatography peaks analyte concentration distributions versus outflow peaks analyte concentration distributions by their mechanism of formations/chromatographic principles [13].

Numerous attempts have been made to derive numerical model for the separation process of certain type chromatography, such as thin layer chromatography (TLC) [14], column chromatography [15], gas chromatography (GC) [16], high performance liquid chromatography (HPLC)[17], and gel permeation chromatography (GPC) [18]; however there is currently no universally-accepted underlying mathematical model for the separation process that governs all types of chromatography [19]. Yang et al. first postulated a numerical model for counter current chromatography (CCC) that distinguished the on-chromatography analyte concentration distributions as governed by the mathematical formula for the binomial distribution and outflow analyte concentration distributions as the mathematical formula of the negative binomial distribution, and they came to this by enumerating the first several stages in a table [20]. They then extrapolated their results as the basis of their theory (theory of counter current extraction table, TCCET) [20]. However, their postulations of the formula for the binomial distribution for on-chromatography analyte concentration distributions and the formula for the negative binomial distribution for outflow analyte concentration distributions were based on tabulated enumeration of only the first several stages—they were not rigorously proved for all stages. Moreover TCCET limited to only counter current chromatography.

In this paper, the first generalized theorem to model the analyte concentration distributions for all types of chromatography is proposed using chromatography tables, but we also provide a proof of this theorem using mathematical induction. This proposed theorem states that, for any number of theoretical plates, the on-chromatography analyte concentration distributions exhibit the same formula as binomial distribution, and outflow analyte concentration distributions exhibit the same formula as negative binomial distribution. In both case, we assume diffusion is negligible. We also distinguished the interpretations of the components of the on-chromatography analyte concentration distributions from the interpretations of the random variable and parameters that define the commonly known binomial distribution and share the mathematical formula. Similarly for the outflow analyte concentration distribution we compared the interpretations to the random variable and parameters for the commonly known negative binomial distribution from statistics that shares this mathematical formula. This helps to reduce confusion between the differing interpretations and ramifications of these shared mathematical formulas.

2.2 Method

2.2.1 Model Description

2.2.1.1 The Principles of Chromatography Separation

Mobile phase and stationary phase are the two phases in chromatography. The mobile phase is the phase that moves in a direction toward outlet. It is usually an eluent solution of analytes (substances to be separated). The stationary phase is the materials fixed in position

during the separation process for the chromatography. The analytes are partitioned in both mobile phase and stationary phase governed by specific partition coefficients. The separation occurs when eluent flows in mobile phase carrying analytes down the stream and stationary phase traps a portion of substances, which keeps them temporarily at fixed position. The substances that partition more in mobile phase travel faster than those that partition less in mobile phase. To assist the mathematical derivation of the model, the entire chromatography path (usually column) is divided into discrete sections (known as theoretical plate) in which the partition in which partitioning of the solute between the stationary phase and the mobile phase is assumed to reach equilibrium. In practice, the total chromatography separation process is divided into n discrete stages, and each stage solvent front travels a distance of stage length and stage length is equal to theoretical plate length.

2.2.1.2 Model Assumptions

(1) We assume that the ratio of solute in mobile phase versus in stationary phase is same throughout the chromatography process for the same compound, i.e., the partition coefficient remains constant throughout chromatography process. (2) Equilibrium of partitioning is assumed for each stage of chromatography process. (3) We assume that diffusion of solute is negligible. (4) We assume that the eluent flow rate is constant throughout the chromatography process. (5) The initial feeding the analytes containing solution occurs only in stage 1, and after that the only intake is the pure eluent. (6) For convenience and without loss generality, we normalized the total amount of analyte that initially had been added to system as the dimensionless quantity 1 to represent the total weight the analyte. Note that this assumption does not change the shape of the chromatography analyte concentration distribution, and that it is the distributions relative shape

that is essential in parameter estimation. Normalization is a technique used for quantitatively assessing a chromatography analyte concentration distribution to provide a quantitative analysis of the mixture being separated.

2.2.2 The Chromatography Process

In order to visualize the chromatography process, we used figure 1 to assist description of the mechanism of chromatography analyte concentration distribution of single analyte. We define partition coefficient ρ as the ratio of the concentration of an analyte in mobile phase to the concentration of this analyte in stationary phase at equilibrium. We also denote the proportion constant λ as:

$$\lambda = \frac{\rho}{1+\rho}$$

It is the proportion of the analyte in mobile phase at equilibrium.

2.2.2.1 The Flow Chart

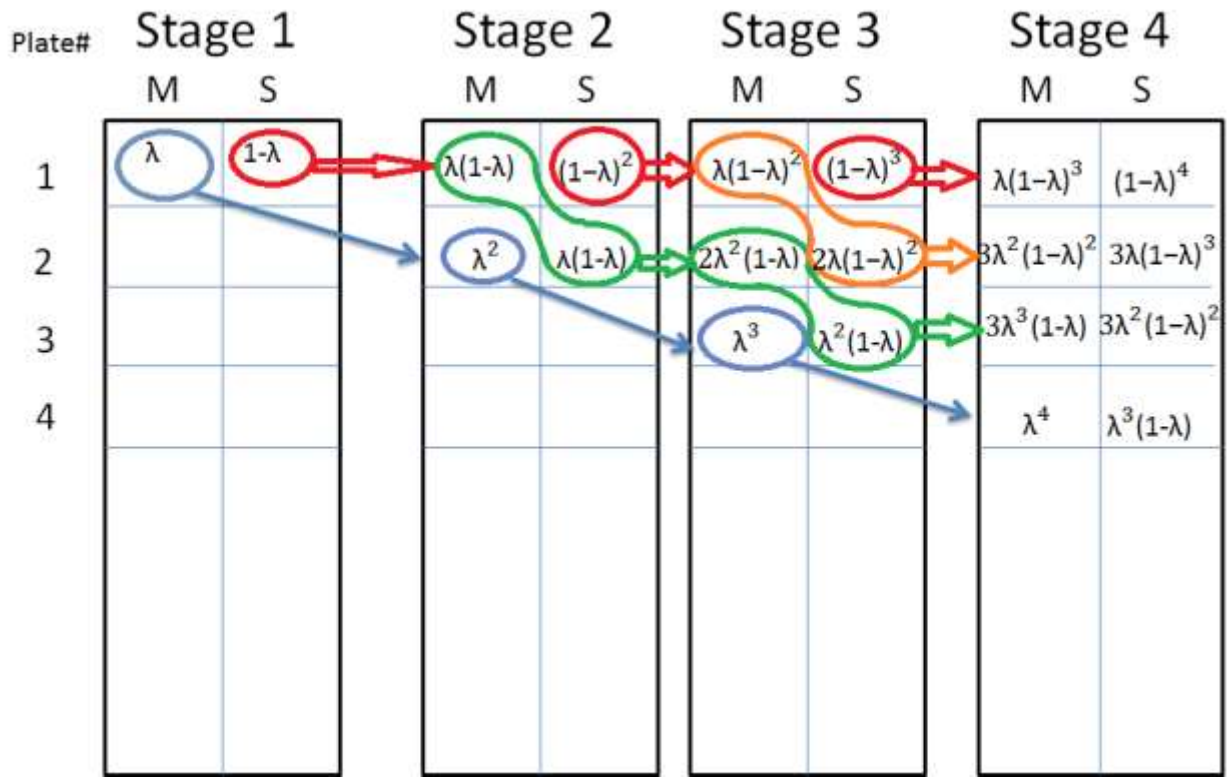
At initial stage (stage 1), we assume that analytes containing solution fed to chromatography is partitioned to both mobile and stationary phase and established equilibrium so that mobile phase contains λ portion of total analyte whereas the stationary phase contains $1-\lambda$ portion of total analyte. In this stage, the frontier of the analyte containing solution reaches the length of one theoretical plate thus as shown in figure 1, analyte is only contained in first plate of chromatography, with λ portion in mobile phase and $1-\lambda$ portion in stationary phase.

At stage 2, in plate 1 of the chromatography the mobile phase is replaced by blank eluent and the mobile phase from stage 1 flows from plate 1 to plate 2. Since the stationary phase temporarily fixed the analyte at the plate 1 from stage 1, plate 1 contains $1-\lambda$ portion of the total analyte. And since there is no analyte in plate 2 of stationary phase from the stage 1, at stage 2 the plate 2 contains λ portion of the total analyte. Once the partition reach equilibrium for this stage (2), $1-\lambda$ portion of the analyte remaining in all plates is partitioned to the stationary phase and λ portion of the analyte in all plates is partitioned to the mobile phase. Hence at stage 2: in the mobile phase in plate 1 there is $(1-\lambda)\lambda$ potion of total analyte, and in the stationary phase of plate 1 there is $(1-\lambda)(1-\lambda)$ portion of total analyte; whereas in the plate 2 mobile phase there is $\lambda\lambda$ portion of total analyte, and in the plate 2 stationary phase there is $\lambda(1-\lambda)$ portion of total analyte.

At stage 3, the mobile phase in plate1 is again replaced by blank eluent, and the mobile phase from stage 1 flows from plate 1 to plate 2, and from plate 2 to plate 3. Since the stationary phase temporarily fixed the analyte at plate 1 from stage 2, plate 1 contains $(1-\lambda)^2$ portion of the total analyte. Plate 2 contains analytes from mobile phase of the plate 1 in stage 2 and the stationary phase of plate 2 in stage 2, and both combined are $(1-\lambda)\lambda$ portion of total analyte. Therefore, in stage 3 plate 2 contains $2(1-\lambda)\lambda$ portion of the total analytes. The stationary phase of plate 3 does not contain any analyte at stage 2, thus at stage 3 all analyte contained in plate 3 are from the mobile phase of plate 2 in previous stage, which is λ^2 portion of total analyte. After the partition reach equilibrium, $1-\lambda$ portion of the analyte in all plates is partitioned to the stationary phase, and λ portion of the analyte in all plates is partitioned to the mobile phase. At this stage, there are $(1-\lambda)^2\lambda$ portion of total analyte in the mobile phase of plate 1, $(1-\lambda)^2(1-\lambda)$ portion of total analyte in the stationary phase of plate 1, $2(1-\lambda)\lambda\lambda$

portion of total analyte in the mobile phase of plate 2, $2(1-\lambda)\lambda(1-\lambda)$ portion of the total analytes in stationary phase of plate 2, $\lambda^2\lambda$ portion of the total analytes in mobile phase of plate 3, and $\lambda^2(1-\lambda)$ portion of the total analytes in the stationary phase of plate 3.

Figure 2.1 Flow chart of analyte concentration in chromatography



This process continues in same manner for each stage, i.e., blank eluent replaces the solution in plate 1, and the mobile phase in each plate from the previous stage carries analyte in the mobile phase to the next plate, whereas the stationary phase temporarily holds the anaylte in same plate as the previous stage. Once the equilibrium is reached, the ratio of anaylte in mobile phase to anaylte in stationary phase in any plate of chromatography is same as the partition coefficient ρ, so $\frac{\rho}{1+\rho}$ portion of analyte is in mobile phase and $\frac{1}{1+\rho}$ portion of analyte is in stationary phase. Thus the recursion relationship for the chromatography can be described as

equation below. The proportion of analytes in mobile phase, stationary phase and combined phases for each plate of chromatography at first four stages are summarized in table 2.1 (similar to [12]).

2.2.2.2 The Chromatography Analyte Concentration Distribution Derived Using Chromatography Table.

Table 2.1 Solute analyte concentration distribution on-chromatography with four theoretical plates.

Plate	Stage 1	Stage 2	Stage 3	Stage 4
	Mobile phase			
1	λ	$\lambda(1-\lambda)$	$\lambda(1-\lambda)^2$	$\lambda(1-\lambda)^3$
2		λ^2	$2\lambda^2(1-\lambda)$	$3\lambda^2(1-\lambda)^2$
3			λ^3	$3\lambda^3(1-\lambda)$
4				λ^4
Sum	λ	λ	λ	λ
	Stationary phase			
1	$1-\lambda$	$(1-\lambda)^2$	$(1-\lambda)^3$	$(1-\lambda)^4$
2		$(1-\lambda)\lambda$	$2\lambda(1-\lambda)^2$	$3\lambda(1-\lambda)^3$
3			$\lambda^2(1-\lambda)$	$3\lambda^2(1-\lambda)^2$
4				$\lambda^3(1-\lambda)$
Sum	$1-\lambda$	$1-\lambda$	$1-\lambda$	$1-\lambda$
	Combined			
1	1	$(1-\lambda)$	$(1-\lambda)^2$	$(1-\lambda)^3$
2		λ	$2(1-\lambda)\lambda$	$3\lambda(1-\lambda)^2$
3			λ^2	$3\lambda^2(1-\lambda)$
4				λ^3
Sum	1	1	1	1

For these stages, the combined proportion of analytes in both mobile phase and stationary phase is distributed with same formula as binomial distribution as shown in Figure 2.2. Assume that we set a cutoff to begin outflow collection at the plate number 2, as shown in figure 2.3. The proportion of analytes in the mobile phase of this second plate from the previous stage represents the proportion of analytes that flow out in the current stage. For example, in stage 2 the outflow

is from the plate 2 mobile phase from stage 1 which is 0; and in stage 4 the outflow is the plate 2 mobile phase from stage 3, which is $2\lambda^2(1-\lambda)$. The proportion of analytes at the flow-out point after passing 2 plates is distributed with same mathematical formula as negative binomial distribution. Generally, for any cut off plate r, in the stage x the outflow is from the plate x-1 mobile phase which is $\binom{x-1}{r-1}\lambda^r(1-\lambda)^{x-r}$.

Figure 2.2 On-chromatography analyte concentration distribution

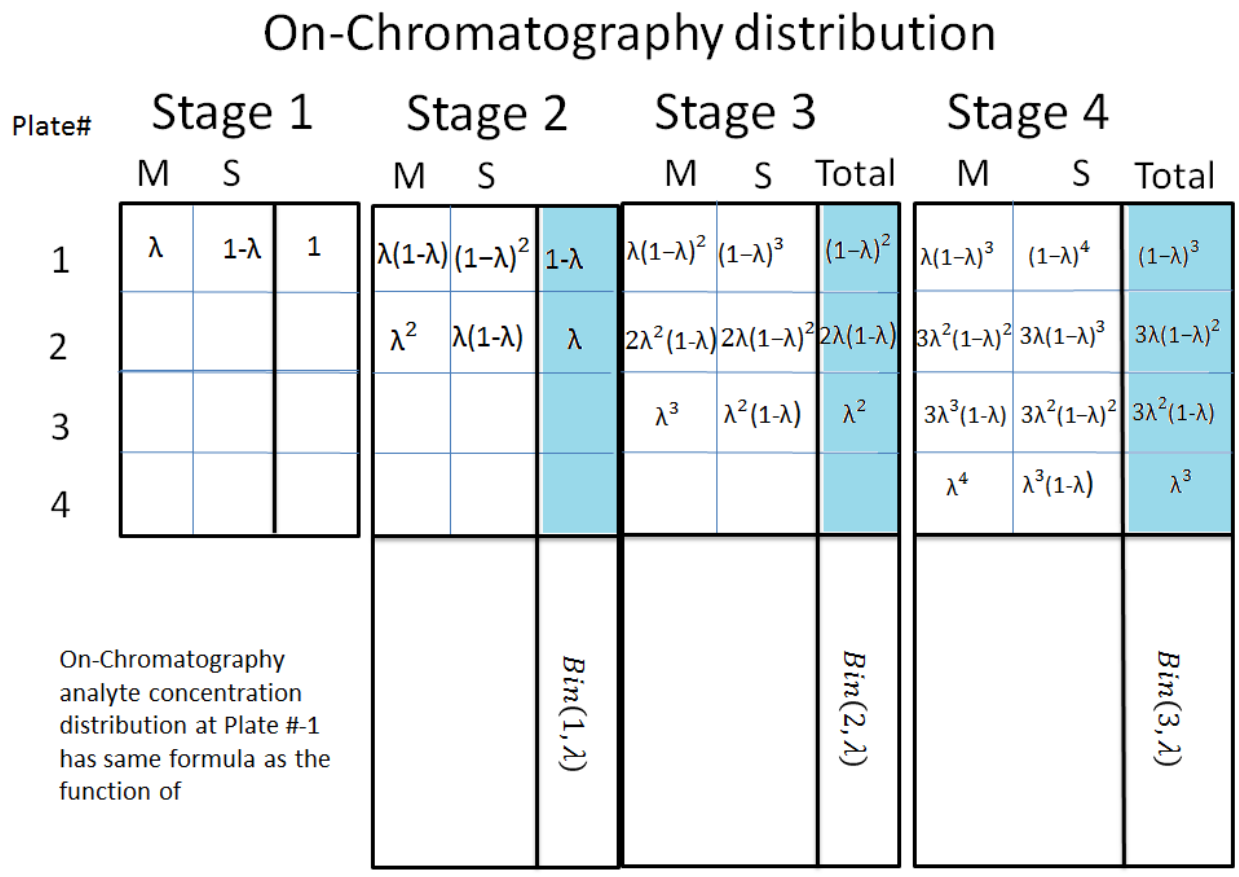
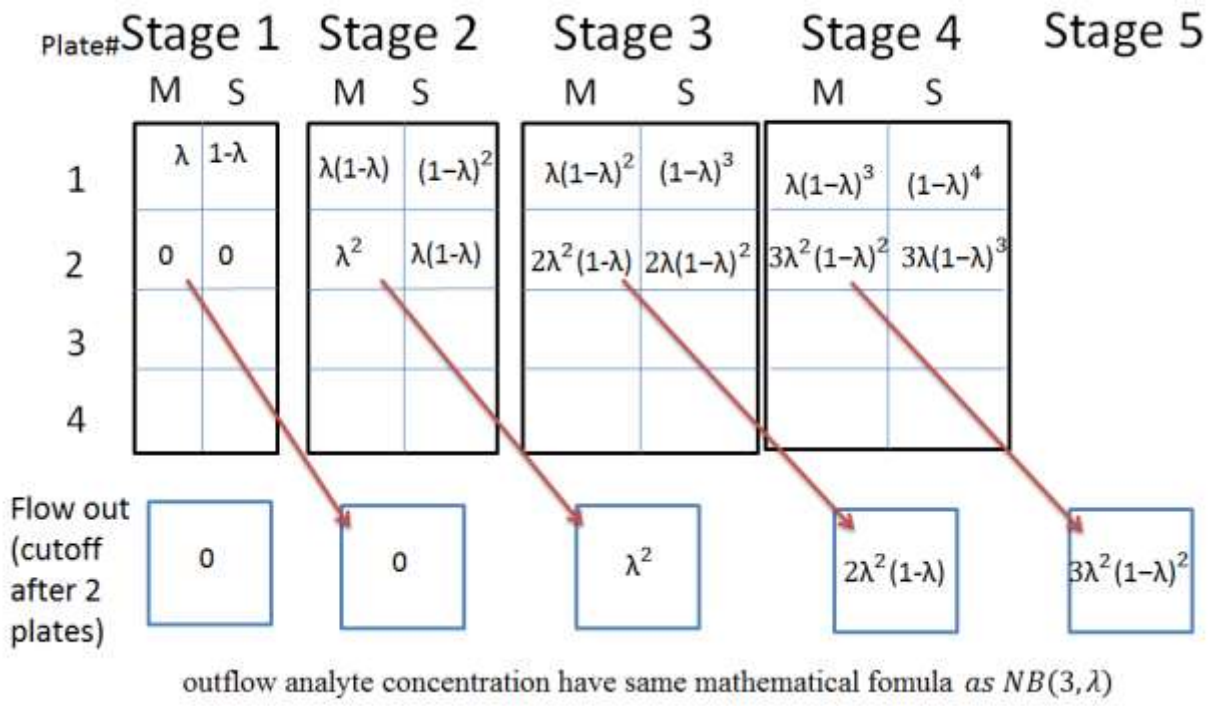


Figure 2.3 Outflow analyte concentration distribution



outflow analyte concentration have same mathematical fomula $as\ NB(3,\lambda)$

We have derived on-chromatography and outflow analyte concentration distribution for initial stages first by tabulation similar to Yang's approach [12]. We then extend Yang's approach [12] to infinite stages for both on-chromatography and flow-out distributions using proof by induction.

2.3. Results

By mathematical induction we prove the formula for the on-chromatography analyte concentration distributions matches the formula for the binomial distribution. Similarly, using induction we prove the formula for outflow analyte concentration distributions match the formula for the negative binomial distribution. These proofs follow.

2.3.1 Notation

For M_i^j , M denotes the relative amount of analyte in mobile phase, the subscript i denotes the stage number, the superscript j denotes the number of theoretical plate.

Similarly, for S_i^j, S denotes the relative amount of analyte in stationary phase, the subscript i denotes the stage number, and the superscript j denotes the number of theoretical plate.

2.3.2 Proof of on-Chromatography and Outflow Analyte Concentration Distributions

2.3.2.1 Initial Condition

Equation (1) describes the first stage of the chromatography process when the analyte was added to the column.

$$\begin{cases} M_1^1 = \lambda \\ S_1^1 = 1 - \lambda \end{cases} \quad (1)$$

2.3.2.2 Recurrence Relation:

Following the initial conditions, blank eluent is added to the column. Equation (2) represents the change in the entire distribution of the analyte concentration along the chromatography (for the specific stage, stage $i + 1$) in comparison to the analyte concentration distribution in the previous stage (stage i) as the equilibrium between the mobile and stationary phase are established within each plate along the chromatography.

$$\begin{cases} M_{i+1}^{j+1} = \left(M_i^j + S_i^{j+1}\right)\lambda \\ S_{i+1}^{j+1} = \left(M_i^j + S_i^{j+1}\right)(1-\lambda) \end{cases} \quad where\ i \in \{1, 2, \dots\}, j \in \{1, \dots, i-1\} \quad (2)$$

2.3.2.3 Proof by Induction.

Next we prove that the on-chromatography analyte concentration distribution as following:

Part (a) proof for the plate (superscript) $\in \{2, \dots i+1\}$

$$\begin{cases} M_i^j = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} \\ S_i^j = \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j+1} \end{cases}$$

First let us prove that it is true for the stage 1:

$$M_1^1 = \binom{1-1}{1-1}\lambda^1(1-\lambda)^{1-1} = \lambda$$

$$S_1^1 = \binom{1-1}{1-1}\lambda^{1-1}(1-\lambda)^{1-1+1} = 1-\lambda$$

$$Therefore\ the\ equation \begin{cases} M_i^j = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} \\ S_i^j = \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j+1} \end{cases} is\ true\ for\ stage\ 1$$

$$Assume\ that\ the\ equation \begin{cases} M_i^j = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} \\ S_i^j = \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j+1} \end{cases} is\ true\ for\ stage\ i,$$

$$plate\ number\ j \in \{1, \dots i\}$$

Let us show that for stage $i+1$

$$\begin{cases} M_{i+1}^{j+1} = \binom{(i+1)-1}{(j+1)-1}\lambda^{(j+1)}(1-\lambda)^{(i+1)-(j+1)} \\ S_{i+1}^{j+1} = \binom{(i+1)-1}{(j+1)-1}\lambda^{(j+1)-1}(1-\lambda)^{(i+1)-(j+1)+1} \end{cases} \quad j \in \{1, \dots, i+1-1\}$$

Which is,

$$\begin{cases} M_{i+1}^{j+1} = \binom{i}{j}\lambda^{j+1}(1-\lambda)^{i-j} \\ S_{i+1}^{j+1} = \binom{i}{j}\lambda^{j}(1-\lambda)^{i-j+1} \end{cases} \quad j \in \{1, \dots, i\}$$

Recall the recursion relation:

$$\begin{cases} M_{i+1}^{j+1} = \left(M_i^j + S_i^{j+1}\right)\lambda \\ S_{i+1}^{j+1} = \left(M_i^j + S_i^{j+1}\right)(1-\lambda) \end{cases}$$

For stage i we assumed that $M_i^j = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j}$ and $S_i^j = \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j+1}$

$\forall\, j \in \{1, \dots, i\}$

then,

$S_i^{j+1} = \binom{i-1}{(j+1)-1}\lambda^{(j+1)-1}(1-\lambda)^{i-(j+1)+1}$, so

$$M_i^j + S_i^{j+1} = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} + \binom{i-1}{(j+1)-1}\lambda^{(j+1)-1}(1-\lambda)^{i-(j+1)+1}$$

$$= \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} + \binom{i-1}{j}\lambda^j(1-\lambda)^{i-j}$$

$$= \left[\binom{i-1}{j-1} + \binom{i-1}{j}\right]\lambda^j(1-\lambda)^{i-j}$$

$$= \left\{\frac{(i-1)!}{[(i-1)-(j-1)]!\,(j-1)!} + \frac{(i-1)!}{[(i-1)-j]!\,j!}\right\}\lambda^j(1-\lambda)^{i-j}$$

$$= \left[\frac{j(i-1)!}{j(i-j)!\,(j-1)!} + \frac{(i-j)(i-1)!}{(i-j)(i-1-j)!\,j!}\right]\lambda^j(1-\lambda)^{i-j}$$

$$= \frac{(i-j+j)(i-1)!}{(i-j)!\,j!}\lambda^j(1-\lambda)^{i-j}$$

$$= \frac{i!}{(i-j)!\,j!}\lambda^j(1-\lambda)^{i-j}$$

$$= \binom{i}{j}\lambda^j(1-\lambda)^{i-j} \quad (3)$$

Therefore, by recursion relation (2) for mobile phase we have:

$$M_{i+1}^{j+1} = \left(M_i^j + S_i^{j+1}\right)\lambda = \binom{i}{j}\lambda^j(1-\lambda)^{i-j}\lambda = \binom{i}{j}\lambda^{j+1}(1-\lambda)^{i-j} \quad (4)$$

By recursion relation (2), for stationary phase we have:

$$S_{i+1}^{j+1} = \left(M_i^j + S_i^{j+1}\right)(1-\lambda) = \binom{i}{j}\lambda^j(1-\lambda)^{i-j}(1-\lambda) = \binom{i}{j}\lambda^j(1-\lambda)^{i-j+1} \quad (5)$$

So for stage $i+1$, $\begin{cases} M_{i+1}^{j+1} = \binom{i}{j}\lambda^{j+1}(1-\lambda)^{i-j} \\ S_{i+1}^{j+1} = \binom{i}{j}\lambda^j(1-\lambda)^{i-j+1} \end{cases}$ is proved for plate $j \in \{1, \ldots i\}$

And thus plate $j+1 \in \{2, \ldots i+1\}$

Part (b) proof for the plate (subscript) $\in \{1\}$

The analyte concentration for mobile and stationary phases of the first plate M_i^1 and S_i^1 $i = 2, 3, ...$

The recursion relation is:

$$M_i^1 = (M_{i-1}^0 + S_{i-1}^1)\lambda = S_{i-1}^1\lambda$$

$$S_i^1 = (M_{i-1}^0 + S_{i-1}^1)\lambda = S_{i-1}^1(1-\lambda)$$

Since after initial addition of analyte only pure eluent was flushed through the system, we have

$$M_{i-1}^0 = 0 \; \forall \; i = 2, 3, ...$$

From initial condition:

$$S_1^1 = 1 - \lambda$$

And $S_i^1 = S_{i-1}^1(1-\lambda)$

Thus $S_i^1 = (1-\lambda)^i$ and $M_i^1 = S_{i-1}^1\lambda = \lambda(1-\lambda)^{i-1}$

Combining the results from part (a) and part (b) we have:

For stage $i+1$, $\begin{cases} M_{i+1}^{j+1} = \binom{i}{j}\lambda^{j+1}(1-\lambda)^{i-j} \\ S_{i+1}^{j+1} = \binom{i}{j}\lambda^{j}(1-\lambda)^{i-j+1} \end{cases}$ is proved for plate $j \in \{1, ... i+1\}$

And hence for stage i $\begin{cases} M_i^j = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} \\ S_i^j = \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j+1} \end{cases}$ $j \in \{1, \dots i\}$ (6)

Denote T_i^j as the total analytes in j[th] plate of i[th] stage then,

$$T_i^j = \left(M_i^j + S_i^j\right) = \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j} \; \forall \; i \; \in \{1,2,3,\dots\} \; and \; j \in \{1,\dots,i\} \quad (7)$$

which has the same mathematical formula as the binomial distribution. Therefore it is proved that the on-chromatography analyte concentration is distributed by the mathematical formula of the binomial distribution for all stages.

$$M_i^j = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} \; \forall \; i \; \in \{1,2,3,\dots\} \; and \; j \in \{1,\dots,i\} \quad (8)$$

which has same mathematical formula as the negative binomial distribution. Therefore it is proved that the outflow analyte concentration distribution is distributed by the mathematical formula of the negative binomial distribution for all stages. Note: the outflow analyte quantity of current stage is analyte quantity of the mobile phase of the last plate of the chromatography column.

2.3.3 Statistical Distribution Functions versus Chromatography Analyte Concentration Distributions

In statistics, the binomial distribution was used to represent a sequence of independent Bernoulli trials with fixed success rate of p. Although the on chromatography distribution of analyte concentration exhibits the same formula as statistical binomial distribution, it bears

different meaning. The chromatography analyte concentration distribution is a result of the sequence of partitions of analytes in each plate of the chromatography with a particular partition coefficient, ρ. The result of Bernoulli trials is either "success" or "failure," whereas the partition of analytes occurs in proportions. The proportion of analyte in mobile phase $\lambda = \frac{\rho}{1+\rho}$ only after equilibrium is established. Therefore we would like to summarize the comparisons between statistical distribution versus chromatography analyte concentration distributions in table 2.

Table 2.2. Contrast between statistical and chromatography terms.

Parameter/Formula	Terminology	
	Statistics	Chromotography
i	Number of independent Bernoulli trials (often denoted as n)	Stage number, and also the total number of plates in a given stage
λ	"Success" probability of a each of the single independent Bernoulli trials; for each Bernoulli trial, the sample space for possible outcomes is the set {"Success", "Failure"}	Proportion of the analyte in the mobile phase at its equilibrium; more specifically, this is: 1) the proportion that can be found by summing the proportion of the analyte in the mobile phase across the chromatography; and 2) the proportion of the analyte in the mobile phase within a given plate (within a given stage)
j	Number of "successes" across the i (n) independent Bernoulli trials (often denoted as Y)	Location, or plate number, on the chromatography; this is also the number that, using the

		binomial distribution formula, allows one to determine how the proportion of the analyte is distributed across the choromotography (this is the proportion in both the mobile and stationary phases at that location/plate)
$\binom{i}{j}\lambda^j(1-\lambda)^{i-j}$	Probability mass function of the number of "successful" independent Bernoulli trials (j, which is often denoted as Y) across the total number of these trials (i, which is often denoted as n)	The proportion of the analyte on the chromatography at plate j in stage i

2.4. Simulations

We simulated the outflow analyte concentration distributions from a chromatography process by using the initial condition (eq. 1) and recursion relationships (eq. 2). We also assume that the random measurement error follows normal distribution with mean 0 and (small) variance 0.0009 (standard deviation of 0.03). The simulated analyte concentration distribution is compared with negative binomial distribution and Gaussian distribution in different scenarios listed in Figure 2.4. For each scenario, all three distributions are exhibited in same plot.

The distribution function of negative binomial distribution is:

$$f(x) = \binom{x-1}{j-1}\lambda^j(1-\lambda)^{x-j}$$

Where $\lambda=\frac{\rho}{1+\rho}$ and ρ is the partition coefficient, j is the total number of theoretical plates and n is the total number of stages.

Gaussian model is obtained by matching the mode and variance with negative binomial model.

$$Mode = \frac{(1-\lambda)(j-1)}{\lambda}$$

$$Variance = \frac{(1-\lambda)j}{\lambda^2}$$

The distribution function of Gaussian model is:

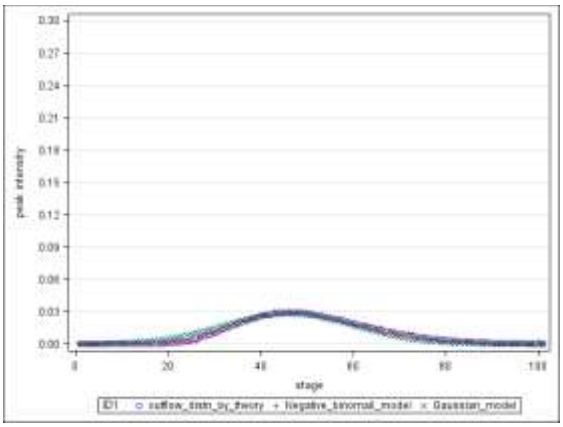
$$f(x) = \frac{1}{\sqrt{2\pi \frac{(1-\lambda)j}{\lambda^2}}} \exp\left(-\frac{\left(x - \frac{(1-\lambda)(j-1)}{\lambda}\right)^2}{2\frac{(1-\lambda)j}{\lambda^2}} \right)$$

The simulation program is develop using a statistical software SAS and for detailed code see appendix.

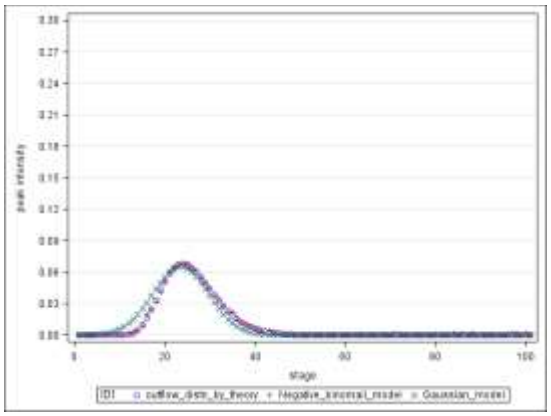
As shown in figure 2.4, the simulated analyte concentration distribution is closer to the negative binomial distribution in that both are right skewed with a heavy tail, whereas the Gaussian model peak is symmetric. If we use the Gaussian model to predict outflow the chromatography peaks in separation process, we may misestimate the cutoff point for collecting one of the components, and thus result in impurities in the later component. For example, assume that λ=0.8, j=5, n=50, as shown in figure 2.5, if we specify a 95% desired recovery, then the cutoff should be at stage 18 by using Gaussian model, however, the real recovery by simulation is 87%, similar to the negative binomial distribution. By using Gaussian model, we would assume that 5% of this analyte will be carried over to the next analyte component, however it is underestimated by 160%. In fact, accord to simulation 13% of this analyte will be carried over to the next analyte component.

Figure 2.4 Simulated outflow chromatography analyte concentration distributions

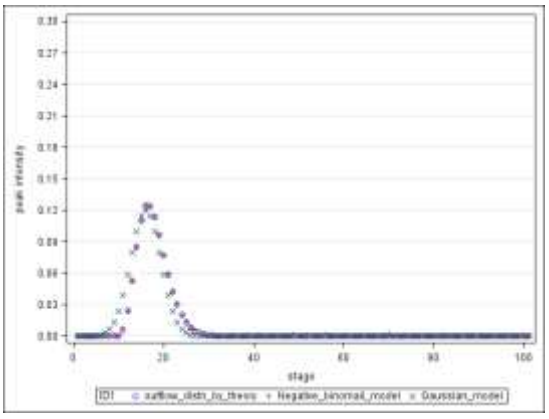
λ=0.2, j=10, n=100



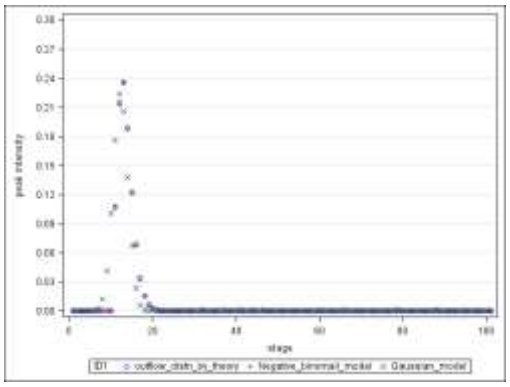
λ=0.4, j=10, n=100



λ=0.6, j=10, n=100

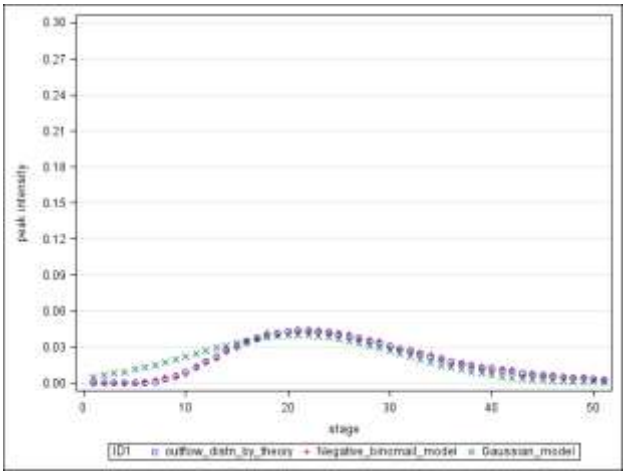


λ=0.8, j=10, n=100

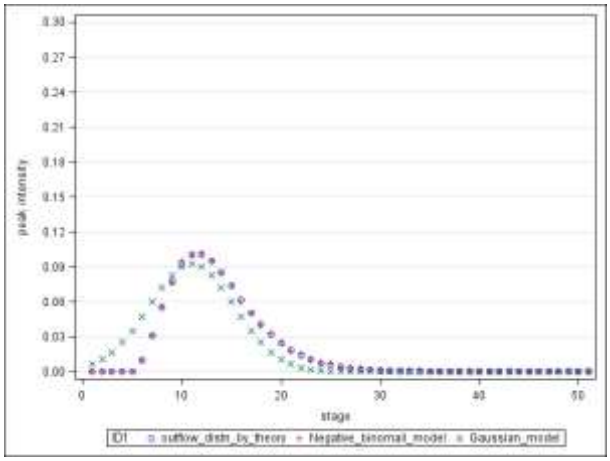


λ=0.2, j=5, n=50

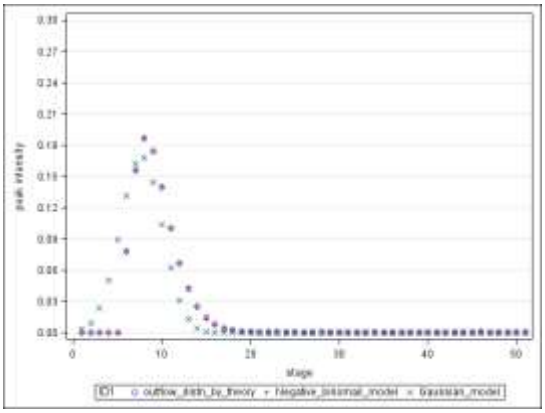
λ=0.4, j=5, n=50



λ=0.6, j=5, n=50



λ=0.8, j=5, n=50



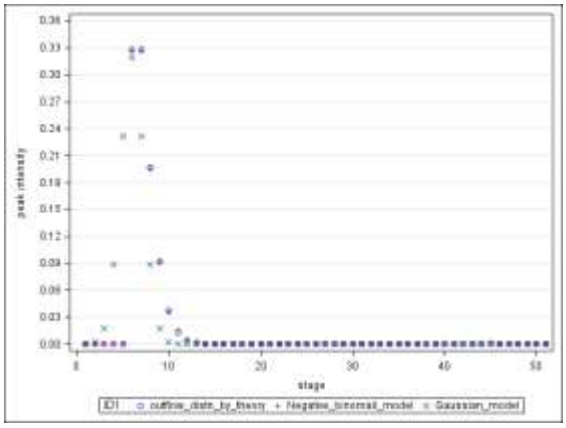
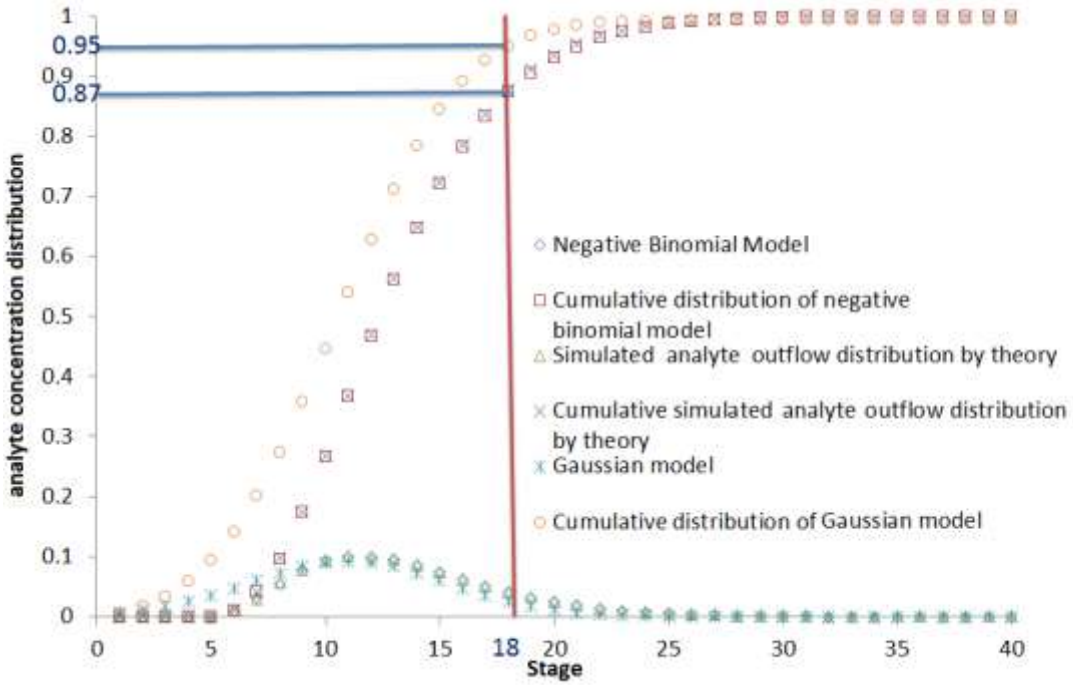


Figure 2.5 Simulated outflow chromatography analyte concentration distributions for λ=0.8, j=5, n=50 and 95% desired recovery.



2.5. Conclusions

In this work we have demonstrated that in general chromatography analyte concentration distributions can be described with formulas that are analogous to formulas for the binomial distribution or the negative binomial distribution. The out-flow analyte concentration distribution formulas are analogous to those of the negative binomial distribution, and the on-chromatography analyte concentration distributions also match the formula for the binomial distribution. Although this result has been tabulated by enumeration of first several stages, there was previously no rigorous proof such that chromatography with infinite stages can be covered. In this study, we not only established the on-chromatography and out-flow analyte concentration

distributions using mathematical proof by induction, but also demonstrated that the simulated outflow analyte concentration distributions match the negative binomial distribution and not the Gaussian distribution, which is often used in practice, by simulations.

Page left intentionally blank

Chapter 3: Binomial-Negative Binomial Theorem, the Mathematic Bridge between the Distribution of the on-Chromatography and the Outflow Analyte Concentration Distributions

3.1. Introduction

The binomial and negative binomial distribution had been of research interest for decades [21]. For example, negative binomial distribution was well known for its fitting to the over-dispersed count data produced by a Poisson mechanism [22]. The binomial distribution was used as the numeric model for chromatography analyte concentration distributions [23]. Recently, the mathematical formula for the negative binomial distribution was shown to be a better numeric model for chromatography analyte concentration distributions in counter current chromatography (CCC) outflow peaks [24]. However, other than under the assumptions of a series of independent Bernoulli trials, the mathematical relation between binomial and negative binomial distribution is still to be explored.

The chromatography is a type of laboratory techniques that separate the mixture of compounds to obtain the compounds of interest. The substance to be separated is called analyte. The chromatography analyte concentration distribution is measured by partitioning of this analyte between mobile and stationary phases. The mobile phase is the phase that moves the mixture in a certain direction and the stationary phase is the phase that fixes, or holds, some of the analyte from the mixture in the place. In chromatography separation the analyte concentration distribution, which is the distribution of the quantity of analyte as it flows out of the chromatography as function of time, provides the theoretical guidance for analyte concentration distribution simulations so that researcher could be able to predict the peak location by using previous experimental data.

Despite a several decades effort to specify the chromatography analyte concentration distribution correctly, the best mathematical model for chromatography separation process is still under debate. For example, Yuri Kalambet *et al* believes that exponentially modified Gaussian function is the best formula for describing chromatographic peak shape [25] however according to F.C. Denizot and M.A. Delaage the chromatography analyte concentration distribution converge toward Laplace-Gaussian distribution[26]. None of these studies account for the difference between the skewness of analyte's outflow and on-chromatography analyte concentration distribution.

In our preceding work, we proved that the on-chromatography analyte concentration distribution is matches that of a binomial distribution, and that the outflow analyte concentration distribution matches that of a negative binomial distribution. This work clarifies the difference and finds the relationship between the on-chromatography analyte concentration distribution and outflow analyte concentration distribution [27]. This relationship between on-chromatography analyte concentration distribution and outflow analyte concentration distribution is important, and is unknown to most researchers in the field of chromatography separation.

Casella and Berger have demonstrated the relationship between binomial and negative binomial cumulative distribution functions, e.g. ex 3.12. [27] They prove this result in their solution manual based on a sequence of Bernoulli trials. In the context of this work, we are unable to rely on the probabilistic relationships that advance their proof. This work relies purely on mathematical relationships between on-chromatography analyte concentration outflow analyte concentration distributions.

3.2. Process of Chromatography Separation

Denote mobile phase's i^{th} stage and j^{th} plate as M_i^j and stationary phase's i^{th} stage and j^{th} plate as S_i^j. Similiarly, denote total analyte in the mixture during the i^{th} stage at the j^{th} plate as T_i^j. We define partition coefficient ρ, as the ratio of the concentration of an analyte in mobile phase to the concentration of this analyte in stationary phase at equilibrium. We also denote the proportion constant λ as:

$$\lambda = \frac{\rho}{1+\rho}$$

It is the proportion of the analyte in the mobile phase of the mixture at equilibrium.

In chapter 2, we proved that

$$\begin{cases} M_i^j = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} \\ S_i^j = \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j+1} \end{cases}$$

$$T_i^j = \left(M_i^j + S_i^j\right) = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} + \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j+1}$$

$$= \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j}[\lambda + 1 - \lambda] = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j}$$

T_{i+1}^{j+1}as function of j represents the total on chromatography analyte concentration as function of (or located at) plate numbers. This analyte concentration distribution has the same formula as the binomial distribution from statistics.

$$f(j) = T_{i+1}^{j+1} = \binom{i}{j} \lambda^j (1-\lambda)^{i-j}$$

When we fix the cutoff at plate number j, the proportion of analyte in mobile phase at last plate across all stages is function of total number of plate i and is the outflow analyte concentration distribution; and this outflow analyte concentration distribution has same as the mathematical formula as negative binomial distribution:

$$f(i) = M_i^j = \binom{i-1}{j-1} \lambda^j (1-\lambda)^{i-j}$$

Assume the k^{th} stage is the stage to complete the dead volume (dead volume is defined as the total volume of the mobile phase in the chromatographic column).

When we fix total number of plates i, at k^{th} stage ($k \leq i$) the total analyte remaining on the chromatography is:

$$\sum_{j=0}^{k-1} \binom{i}{j} \lambda^j (1-\lambda)^{i-j}$$

At k^{th} stage the total analyte in collection of outflow solutions are:

$$\sum_{l=k}^{i} \binom{l-1}{k-1} \lambda^k (1-\lambda)^{l-k}$$

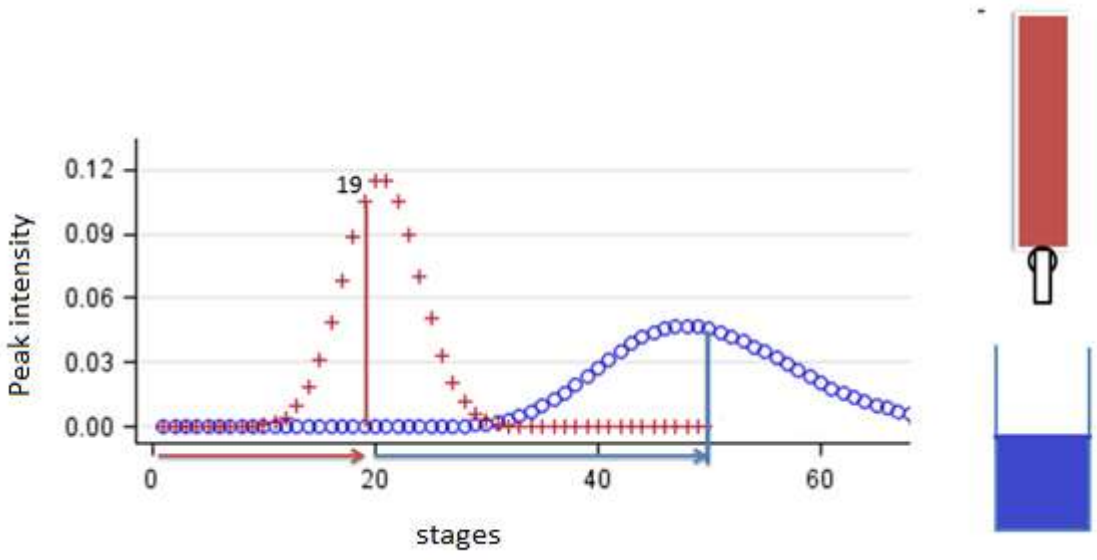
Thus the proposition of our (mathematically-based) binomial-negative binomial theorem in this setting is:

$$\sum_{j=0}^{k-1} \binom{i}{j} \lambda^j (1-\lambda)^{i-j} + \sum_{l=k}^{i} \binom{l-1}{k-1} \lambda^k (1-\lambda)^{l-k} = 1 \qquad (1)$$

This implies that the combination of the sum of the on-chromatography analyte concentration distribution from plate 1 to plate k and the sum of outflow chromatography analyte concentration distribution from the stage k to i is one (1).

We illustrated this relationship (eq 1.) by an example as shown in Figure 3.1. Suppose we set the cutoff for the flow out distribution of analyte at 20th Plate and set the total stage number to be 50. Then, the on chromatography analyte concentration distribution is the sum of analyte concentrations from plate 1 to plate 19, the red crosses in Figure 3.1 and the out-flow analyte concentration distribution is the sum of analyte concentrations in mobile phase of 19th plate from stage 20 to stage 50, the blue circles in Figure 3.1. If we add these two sums together it should equal the total of all or the analyte concentrations, or one (1).

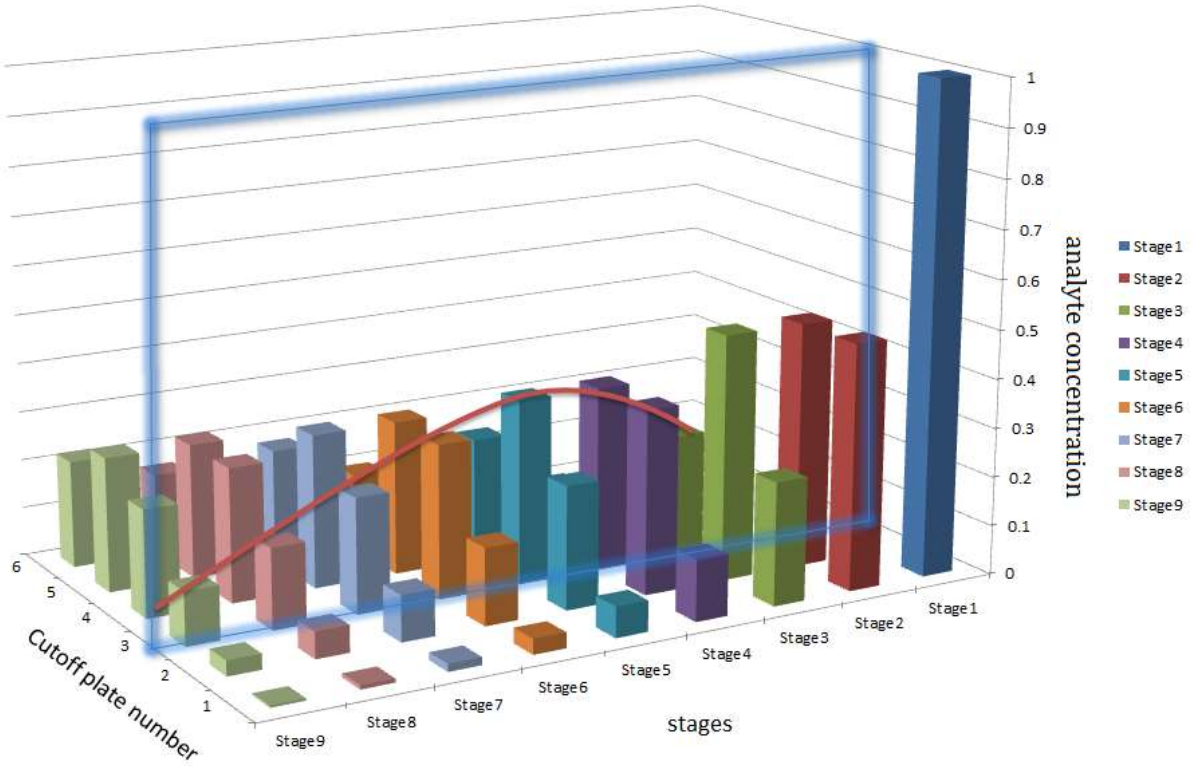
Figure 3.1. Diagram of chromatography process

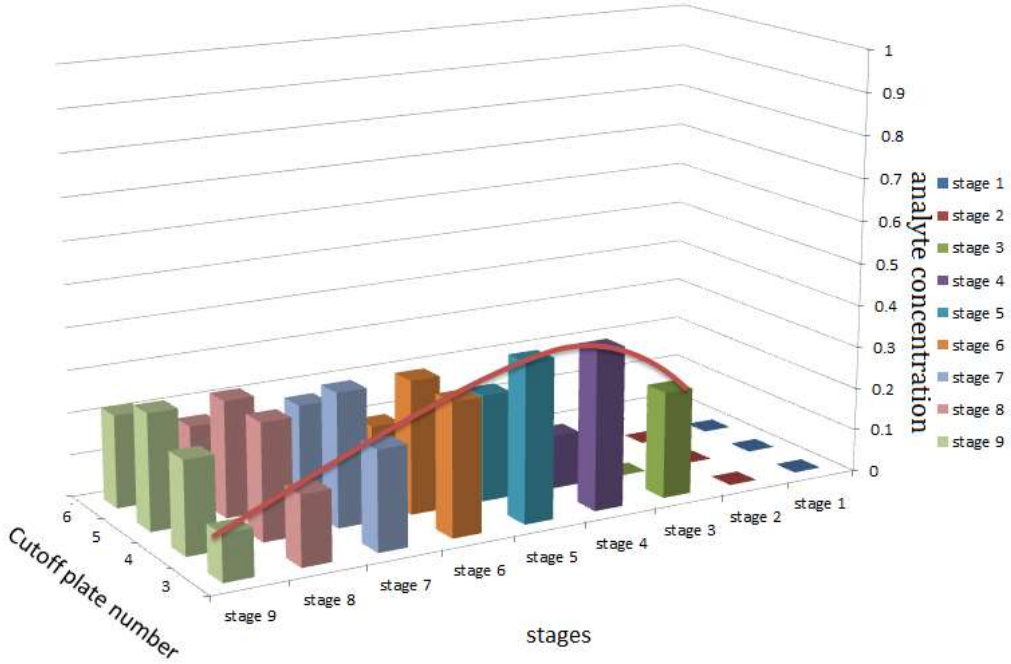


The relation between the on-chromatography analyte concentration distribution (which has been shown in Chapter 2 to match the binomial distribution) and the outflow analyte

concentration distribution (which has been shown in Chapter 2 to match the negative binomial distribution) can be better visualized in figure 3.2, a 3-D plot of how the on-chromatography analyte concentration distribution changes over the first 9 stages, assuming the partition coefficient to be 1 and thus the proportion constant $\lambda = 0.5$. In each stage the quantity of the analyte concentration for each plate is represented by the bars in a particular color, which again, notably, mathematically matches the binomial distribution (indicated by T_i^j above). Suppose we set the cutoff for chromatography at the 3rd plate, and thus the outflow analyte concentration distribution at each stage equals to the concentration of the analyte in last plate multiplied by the proportion constant λ in the previous stage (the analyte concentration in the mobile phase of last plate). Therefore analyte concentration distribution across the highlighted pane multiply the proportion constant λ represents the outflow analyte concentration distribution over the first 9 stages, which matches the mathematical formula for the negative binomial distribution

Figure 3.2 3-D plot of on-chromatography analyte concentration distribution (top). 3-D plot of on-chromatography analyte concentration distribution taking off the first two plates for clarity of display of outflow analyte concentration distribution (bottom)


analyte concentration
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Stage 6
Stage 7
Stage 8
Stage 9
Cutoff plate number
6
5
4
3
2
1
Stage 9
Stage 8
Stage 7
Stage 6
Stage 5
Stage 4
Stage 3
Stage 2
Stage 1
stages


analyte concentration
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
stage 1
stage 2
stage 3
stage 4
stage 5
stage 6
stage 7
stage 8
stage 9
Cutoff plate number
6
5
4
3
stage 9
stage 8
stage 7
stage 6
stage 5
stage 4
stage 3
stage 2
stage 1
stages

The data for the plot of figure 3.2 is listed in table 3.1. The bordered data colored in green and yellow represent a particular example that follows the mathematically-derived binomial-negative binomial theorem $\sum_{j=0}^{k-1}\binom{i}{j}\lambda^j(1-\lambda)^{i-j}+\sum_{l=k}^{i}\binom{l-1}{k-1}\lambda^k(1-\lambda)^{l-k}=1$. In this example, the on-chromatography binomial portion $\sum_{j=0}^{k-1}\binom{i}{j}\lambda^j(1-\lambda)^{i-j}$ are the cells colored in yellow and the outflow negative binomial potion $\sum_{l=k}^{i}\binom{l-1}{k-1}\lambda^k(1-\lambda)^{l-k}$ are the cells colored in green multiply the proportion constant $\lambda = 0.5$. Through addition, it can be shown that $0.015625+0.09375+0.234375+\frac{0.25+0.375+0.375+0.3125}{2}=1$. And if we set the cutoff of the chromatography to be the 4th plate then at stage 8, the on-chromatography portion (that matches the formula for the binomial distribution) $\sum_{j=0}^{k-1}\binom{i}{j}\lambda^j(1-\lambda)^{i-j}$ are the cells colored in pink, and $\sum_{l=k}^{i}\binom{l-1}{k-1}\lambda^k(1-\lambda)^{l-k}$ (the outflow that matches the formula for the negative binomial distribution) are the cells colored in blue multiplied by the proportion constant $\lambda = 0.5$. Summing these quantities again adds to one (1):

0.078125+0.0546875+0.1640625+0.2734375+0.5*(0.125+0.25+0.3125+0.3125)=1.

Table 3.1Analyte concentration at first 9 stages across first 9 plates

Plate #	Stage1	Stage2	Stage3	Stage4	Stage5	Stage6	Stage7	Stage8	Stage9
1	1	0.5	0.25	0.125	0.0625	0.03125	0.015625	0.0078125	0.003906
2		0.5	0.5	0.375	0.25	0.15625	0.09375	0.0546875	0.03125
3			0.25	0.375	0.375	0.3125	0.234375	0.1640625	0.109375
4				0.125	0.25	0.3125	0.3125	0.2734375	0.21875
5					0.0625	0.15625	0.234375	0.2734375	0.273438
6						0.03125	0.09375	0.1640625	0.21875
7							0.015625	0.0546875	0.109375
8								0.0078125	0.03125
9									0.003906

Now if we let $n = i, x = j, y = l, r = k, p = \lambda$. We make this switch to match more commonly used notation for the formulas of the binomial and negative binomial distributions as follows.

Let analyte concentration be a function $f(\cdot)$ of plate number (x) for on-chromatography analyte concentration distribution.

And let analyte concentration be a function $g(\cdot)$ of stage number (y) for the outflow analyte concentration distribution.

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x};\ g(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r}$$

3.3 Proof of Binomial-Negative Binomial Theorem Based Solely on Mathematical Relationships

Proposition (Binomial-Negative Binomial Theorem)

$$\sum_{x=0}^{r-1} \binom{n}{x} p^x (1-p)^{n-x} + \sum_{y=r}^{n} \binom{y-1}{r-1} p^r (1-p)^{y-r} = 1 \qquad (2)$$

Before prove this Proposition we need to prove the following Lemma:

$For\ p \in (0,1), r \in \aleph, n = r+1, r+2, \ldots$

$Let\ \varphi(n)\ be\ a\ function\ of\ n, defined\ as\ below$:

$$\varphi(n) = \binom{n}{r} p^r (1-p)^{n-r} + \sum_{y=r+1}^{n} \binom{y-1}{(r+1)-1} p^{r+1} (1-p)^{y-(r+1)}$$

$$- \sum_{y=r}^{n} \binom{y-1}{r-1} p^r (1-p)^{y-r}$$

$Then\ \varphi(n) = 0 \quad (3)$

$We\ can\ prove\ that\ lemma\ (3)\ is\ true\ by\ induction\ as\ following$:

$For\ \ n = r + 1$

$$\varphi(r+1) = \binom{r+1}{r} p^r (1-p)^{(r+1)-r} - \sum_{y=r}^{r+1} \binom{y-1}{r-1} p^r (1-p)^{y-r}$$

$$+ \sum_{y=r+1}^{r+1} \binom{y-1}{(r+1)-1} p^{r+1} (1-p)^{y-(r+1)}$$

$$= (r+1)p^r(1-p) - \binom{r-1}{r-1} p^r (1-p)^0 - \binom{r}{r-1} p^r (1-p)^{r+1-r}$$

$$+ \binom{(r+1)-1}{(r+1)-1} p^{r+1} (1-p)^{(r+1)-(r+1)}$$

$$= (r+1)p^r(1-p) - p^r - rp^r(1-p) + p^{r+1}$$

$$= 0$$

$Therefore\ \varphi(n = r+1) = 0$

$Now\ let\ us\ assume\ that\ equation\ (3)\ hold\ for\ n-1\ (\varphi(n-1) = 0)$

$$\binom{n-1}{r}p^r(1-p)^{(n-1)-r} + \sum_{y=r+1}^{n-1}\binom{y-1}{r}p^{r+1}(1-p)^{y-(r+1)} - \sum_{y=r}^{n-1}\binom{y-1}{r-1}p^r(1-p)^{y-r}$$

$$= 0$$

$$\varphi(n) = \binom{n}{r}p^r(1-p)^{n-r} + \sum_{y=r+1}^{n}\binom{y-1}{r}p^{r+1}(1-p)^{y-(r+1)} - \sum_{y=r}^{n}\binom{y-1}{r-1}p^r(1-p)^{y-r}$$

$$= \binom{n-1}{r}p^r(1-p)^{n-r} + \binom{n-1}{r-1}p^r(1-p)^{n-r} + \sum_{y=r+1}^{n}\binom{y-1}{r}p^{r+1}(1-p)^{y-(r+1)}$$

$$- \sum_{y=r}^{n-1}\binom{y-1}{r-1}p^r(1-p)^{y-r} - \binom{n-1}{r-1}p^r(1-p)^{n-r}$$

$$= \binom{n-1}{r}p^r(1-p)^{n-r} + \sum_{y=r+1}^{n}\binom{y-1}{r}p^{r+1}(1-p)^{y-(r+1)} - \sum_{y=r}^{n-1}\binom{y-1}{r-1}p^r(1-p)^{y-r}$$

$$= \binom{n-1}{r}p^r(1-p)^{n-r} + \sum_{y=r+1}^{n}\binom{y-1}{r}p^{r+1}(1-p)^{y-(r+1)} - \sum_{y=r}^{n-1}\binom{y-1}{r-1}p^r(1-p)^{y-r}$$

$$= \binom{n-1}{r}p^r(1-p)^{n-r} + \binom{n-1}{r}p^{r+1}(1-p)^{n-(r+1)} + \sum_{y=r+1}^{n-1}\binom{y-1}{r}p^{r+1}(1-p)^{y-(r+1)}$$

$$- \sum_{y=r}^{n-1}\binom{y-1}{r-1}p^r(1-p)^{y-r}$$

$$= \binom{n-1}{r} p^r (1-p)^{n-r} \left(\frac{1}{1-p}\right) + \sum_{y=r+1}^{n-1} \binom{y-1}{r} p^{r+1} (1-p)^{y-(r+1)}$$

$$- \sum_{y=r}^{n-1} \binom{y-1}{r-1} p^r (1-p)^{y-r}$$

$$= \binom{n-1}{r} p^r (1-p)^{(n-1)-r} + \sum_{y=r+1}^{n-1} \binom{y-1}{r} p^{r+1} (1-p)^{y-(r+1)}$$

$$- \sum_{y=r}^{n-1} \binom{y-1}{r-1} p^r (1-p)^{y-r}$$

$$= 0$$

Thus the lemma (3) $\varphi(n) = 0$ *hold for all* $n = r+1, r+2, r+3 \ldots$

Now let's prove that equation (2) *is true* $\forall r \in \aleph$ *by induction*

For $r = 1$ *equation* (2) *can be expressed as following*:

$$\sum_{x=0}^{0} \binom{n}{x} p^x (1-p)^{n-x} + \sum_{y=1}^{n} \binom{y-1}{0} p^1 (1-p)^{y-1} = (1-p)^n + p\left(\sum_{y=1}^{n} (1-p)^{y-1}\right)$$

$$let\ A = \sum_{y=1}^{n} (1-p)^{y-1}\ then\ A - A(1-p) = 1 - (1-p)^n$$

$$Thus\ (1-p)^n + p\left(\sum_{y=1}^{n} (1-p)^{y-1}\right) = (1-p)^n + Ap = 1$$

$So\ theorem\ hold\ for\ r = 1$

$Now\ assume\ that\ the\ binomial\ negative\ binomial\ theorem\ (2)\ hold\ for\ r$

$$Let\ \Psi(r) = \sum_{x=0}^{r-1} \binom{n}{x} p^x (1-p)^{n-x} + \sum_{y=1}^{n} \binom{y-1}{r-1} p^r (1-p)^{y-r} \quad then\ \ \Psi(r)$$

$$= 1$$

$Let\ us\ show\ that\ this\ \ theorem\ hold\ for\ r+1\ which\ is\ \Psi(r+1) = 1$

$$by\ definition, \Psi(r+1) = \sum_{x=0}^{r+1-1} \binom{n}{x} p^x (1-p)^{n-x} + \sum_{y=1}^{n} \binom{y-1}{r+1-1} p^{r+1} (1-p)^{y-(r+1)}$$

$$= \sum_{x=0}^{r-1} \binom{n}{x} p^x (1-p)^{n-x} + \binom{n}{r} p^r (1-p)^{n-r} + \sum_{y=1}^{n} \binom{y-1}{r} p^{r+1} (1-p)^{y-(r+1)}$$

$By\ lemma\ (2)$

$$\varphi(n) = \binom{n}{r} p^r (1-p)^{n-r} + \sum_{y=r+1}^{n} \binom{y-1}{r} p^{r+1} (1-p)^{y-(r+1)}$$

$$- \sum_{y=r}^{n} \binom{y-1}{r-1} p^r (1-p)^{y-r} = 0$$

$$so, \sum_{y=1}^{n} \binom{y-1}{r} p^{r+1} (1-p)^{y-(r+1)} = \sum_{y=r}^{n} \binom{y-1}{r-1} p^r (1-p)^{y-r} - \binom{n}{r} p^r (1-p)^{n-r}$$

$Thus$

$$\Psi(r+1) = \sum_{y=r}^{n} \binom{y-1}{r-1} p^r (1-p)^{y-r} - \binom{n}{r} p^r (1-p)^{n-r}$$

$$+ \sum_{x=0}^{r-1} \binom{n}{x} p^x (1-p)^{n-x} + \binom{n}{r} p^r (1-p)^{n-r}$$

$$= \sum_{x=0}^{r-1} \binom{n}{x} p^x (1-p)^{n-x} + \sum_{y=r}^{n} \binom{y-1}{r-1} p^r (1-p)^{y-r} + \binom{n}{r} p^r (1-p)^{n-r}$$

$$- \binom{n}{r} p^r (1-p)^{n-r}$$

$$= \sum_{x=0}^{r-1} \binom{n}{x} p^x (1-p)^{n-x} + \sum_{y=r}^{n} \binom{y-1}{r-1} p^r (1-p)^{y-r} = \Psi(r)$$

Based on the assumption that $\Psi(r) = 1$ we have that $\Psi(r+1) = 1 \; for \; all \; r \in \aleph \; and \; n = r+1, r+2, \ldots$ and the Binomial-Negative Binomial Theorem is proved based solely on their mathematical formulas.

3.4 Conclusions

In this work, we have demonstrated how the on-chromatography analyte concentration distribution, which matches the mathematical formula used for the binomial distribution, is related to outflow analyte concentration distribution, which matches the mathematical formula for the negative binomial distribution, by the proportion constant, λ. This is visualized by 3-D plot of an on-chromatography analyte concentration distribution example for the first several stages. Based on this relationship we have proposed and proved the binomial-negative binomial theorem by mathematical induction. It can be applied in establishment of relationship between

the on-chromatography and outflow analyte concentration distributinos in chromatography separation processes. This work enables researcher to visualize chromatography separation process by the two simultaneous analyte concentration distributions on-chromatography (matching the binomial distribution formula mathematically) and the outflow (matching the negative binomial distribution formula mathematically), which further helps to clarify the current misunderstanding of chromatography applications that estimate analyte concentration distributions using the same formula (or analyte concentration distribution) for both on-chromatography and outflow chromatography analyte concentration distributions.

Page left intentionally blank

Chapter 4: Estimating Parameters in the Chromatography Separation Process

4.1 Introduction

Chromatography is a widely applied technique for the separation of various mixtures based on their difference in partition coefficient between mobile and stationary phase. Correct identification of the mathematical model of the chromatographic separation process was an essential preliminary step understanding the mechanism of separation and for prediction of retention time of analytes during this process. Many different numeric models such as exponential, Gaussian (normal), exponential modified Gaussian, Weibull, log-normal were proposed to fit the chromatography analyte concentration distributions [28-32] by empirical peak matching; however, most of these models did not follow the mechanism of chromatography separation. The mathematical formulas for the binomial distribution and negative binomial distribution were postulated as the numeric models for on-chromatography and outflow analyte concentration distributions, respectively, for the counter current chromatography (CCC) by Yang et al using theory of countercurrent extraction table (TCCET)[33]. In chapter 2, we proved that the on-chromatography analyte concentration distribution mathematically matches the formula for the binomial distribution, and that the outflow analyte concentration distribution mathematically matches the formula for the negative binomial distribution, and further that this was the case for all type of chromatography separation processes. Since the chromatography analyte concentration distribution has already been rigorously determined (in Chapter 2), we extend this to provide a means of parameter estimation of chromatography analyte concentration distributions by mapping previous established estimation methodology to the types of data or measures collected by chromatography processes. The estimation of the analyte concentration

distribution parameter leads to the estimation of partition coefficient using chromatographic data. This builds the foundation for many applications such as chromatography analyte concentration distribution simulation, analyte component selection, and deconvolution when there are more than one analyte concentration to be separated [34].

To date, the estimation of partition coefficient has been conducted by either a separate set of experiments [35] or by some optimization algorithm, e.g., the particle swarm optimization (PSO) algorithm in certain types of chromatography such as immobilized metal affinity expanded bed adsorption chromatography [36]. The limitations of these methods are: (1) it is time consuming and costly for conducting a different set of experiments; and (2) the particle swarm optimization (PSO) algorithm is only suitable for a particular type of chromatography (immobilized metal affinity expanded bed adsorption chromatography). In this work, we apply the standard statistical methods, (e.g., the method of moment and maximum likelihood estimation) to estimate the chromatography analyte concentration distribution parameters such as number of theoretic plates and partition coefficient between mobile and stationary phase.

Previously we have developed the mathematical model for the separation process of chromatography using discrete formulas in chapter 2. This is more suitable to model the types of chromatography that have a relatively small number of theoretical plates, such as column chromatography, thin layer chromatography (TLC) and counter current chromatography (CCC) since peaks are discrete. However for the types of chromatography with large number of theoretical plates, such as gas chromatography (GC), high pressure liquid chromatography (HPLC), the analyte concentration distributions are closer to continuous, therefore it is more desired to develop a mathematical model using continuous formulas. As plate height approaches zero and the number of plates approaches infinity, the on-chromatography and outflow discrete

distributions become approximately continuous. We apply Taylor expansion and use moment generating functions to approximate the discrete on-chromatography analyte concentration distributions (which were proven in Chapter 2 to match the binomial distribution formula) and discrete outflow analyte concentration distributions (which were proven in Chapter 2 to match the negative binomial distribution formula) with continuous mathematical formulas, which match the formulas for Gaussian distributions. Notably, most of the current application utilized this distribution to simulate outflow analyte concentration distributions for most types of chromatography.[37]

4.2 Chromatography Data

Chromatography experiments produce two data variables. They can either provide peak intensity and retention time for the chromatography with continuous concentration distributions, or the weight of analyte and the volume of mobile phase (eluent) that has run through the column for the chromatography with discrete analyte concentration distribution. The raw column chromatography data for 1,4-dibutoxylbezene from Bai et al. previous chemical compound separation work [38] is exhibited in table 4.1, in which the weight of analyte and the volume of mobile phase (eluent) were recorded. The weight of analyte was obtained by measurement of the dried analyte from the collection of the chromatography outflow solution. The experimental data produced are analogous to a histogram, and notably histograms can be generated from typical data produced by experiments to be statistically analyzed. Notably, chromatography date collected are generally out of scale on both axes as compared to the relative frequencies of a histogram, and thus the outflow chromatography data need to be transformed to be more similar to data generated from typical statistical analyzed experiment in order to apply the established

parameter estimation processes; and because the chromatography data provide the relative shape for a histogram, this become possible.

In chromatography separation, the estimation of the parameters, such as the partition coefficient and number of theoretical plates to use, are of research interest since it can provide information assisting prediction of the location of analyte concentrations. In this study, we transformed data to convert the chromatography data to statistical data with following 3 steps: (1) adjustment by an offset (i.e., the theoretical plate volume); (2) conversion of the weight to frequency to unfold the data as ordered data; and (3) randomize the ordered data.

Table 4.1 Example of raw data from chromatography separation of 1,4-dibutoxylbezene

Volume(mL)	Weight(mg)	Volume(mL)	Weight(mg)	Volume(mL)	Weight(mg)
2	0	114	21	128	20
4	0	116	30	130	13
⋮	⋮	118	38	132	9
106	0	120	41	134	5
108	1	122	39	136	3
110	5	124	34	138	2
112	12	126	27	140	1

We need to transform volume into plate number by the appropriate offset value. This offset serves to normalize the distribution and also match its domain of the outflow analyte concentration distribution, which was found to have the same formula as the negative binomial distribution. The offset is 2 mL (the volume of theoretical plate) in this example, and thus the data is transformed as shown is table 4.2. The volume divided by the offset produces the

theoretical plate number. The volume of theoretical plate can be calculated by dividing dead volume by the total theoretical plate number.

Table 4.2 Transformed data where volume is divided by the offset

Plate#	Weight(mg)	Plate#	Weight(mg)	Plate#	Weight(mg)
1	0	57	21	64	20
2	0	58	30	65	13
⋮	⋮	59	38	66	9
53	0	60	41	67	5
54	1	61	39	68	3
55	5	62	34	69	2
56	12	63	27	70	1

Without loss of generality, the weight of the analyte is converted to observed frequencies by one count per milligram (see Note below) and the data is transformed to order statistics as shown in table 4.3. There is a point for each milligram weight of analyte in this table

Table 4.3 Converted raw data

$x_{(1)}$	54	$x_{(7)}$	56	$x_{(13)}$	56
$x_{(2)}$	55	$x_{(8)}$	56	$x_{(14)}$	56
$x_{(3)}$	55	$x_{(9)}$	56	$x_{(15)}$	56
$x_{(4)}$	55	$x_{(10)}$	56	$x_{(16)}$	56
$x_{(5)}$	55	$x_{(11)}$	56	$x_{(17)}$	56
$x_{(6)}$	55	$x_{(12)}$	56	$x_{(18)}$	⋮

Note: The conversion of observed frequencies by the choice of scale for weight (e.g., milligrams, grams, etc.) of the analyte is relative. Both method of moments (MOM) and maximum likelihood estimator (MLE) are invariant of conversions of frequency counts to different choices of scaling for the weight of the analyte recovered. In other words, the multiplicative change of count per

unit of scale renders same estimation of parameter because the relative shape of the histogram produced by the data derived from the chromatogram results is constant across different scales (Invariance property of MLE is well known and the proof of invariance property of MOM in weight to frequency conversion, see section 4.4.1)

4.3 Continuity Approximation by Asymptotic Chromatographic Analyte Concentration Distributions

In chapter 2, we have proved that the mathematical model for on-chromatography analyte concentration distributions matches the binomial distribution, and the outflow analyte concentration distribution distribution matches the negative binomial distribition for discrete types of chromatography, such as column chromatography. In this chapter we propose the formula for the Gaussian distribution to approximate both on-chromatography and outflow analtye concentration distributions for the continuous chromatography analyte concentration distribution large number of theoretical plate (when partition coefficients are not near their boundaries of zero (0) and one (1)).

4.3.1 Gaussian Approximation of on-Chromatography Analyte Concentration Distribution

In statistics, the normal distribution has been shown to provide a good approximation for the binomial distribution by the central limit theorem [39]. However, in chromatography, although the on-chromatography and outflow analyte concentration distributions as functions of stage and plate number were proved to have the same formula as statistical binomial and

negative binomial distributions, respectively, the functions of analyte concentration distributions are not probabilistic, but rather are deterministic. Therefore, in this work, we present a purely mathematical proof of the validity of Gaussian approximation of analyte concentration distributions—both on-chromatography and outflow—without relying on probability-based relationships.

Assumptions:

Since partition coefficient ρ is assumed to be not close to 0 or infinity, we have the proportion constant $\lambda = \frac{\rho}{1+\rho}$ not close to 0 or 1 (i.e., range of λ is (0,1)).

Denote T_i^j as the total analytes in j[th] plate of i[th] stage then, from equation (2)

$$T_i^j = \left(M_i^j + S_i^j\right) = \binom{i-1}{j-1}\lambda^j(1-\lambda)^{i-j} + \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j+1}$$

$$= \binom{i-1}{j-1}\lambda^{j-1}(1-\lambda)^{i-j}$$

For fixed total number of theoretical plates i, T_i^j as function of $j-1$ has the same formula as that for the binomial distribution $Bin(i-1,\lambda)$

For simplicity, we switch to commonly used binomial distribution formula parameter notation conventions.

Let $f(x)$ have the same mathematical formula as $Bin(n,p)$, and assume that $n \to \infty$

$$f(x) = \binom{n}{x}p^x(1-p)^{n-x}$$

Let us show that $f(x)$ has the same formula as $N\left(np, np(1-p)\right), n \to \infty$

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

Approximate the factorial to exponential using Stirling's equation

$$ln(n!) = nln(n) - n + \frac{1}{2} ln(2\pi n) + \ln\left[1 + \mathcal{O}\left(\frac{1}{n}\right)\right]$$

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(\left[1 + \mathcal{O}\left(\frac{1}{n}\right)\right]\right)$$

which is same as:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Or with bounds:

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq e\sqrt{n}\left(\frac{n}{e}\right)^n$$

$\mathcal{O}(n)$ is a function, defined as order of n. np is of the order of $\mathcal{O}(n)$and $n(1-p)$ is also of the order of $\mathcal{O}(n)$

As $n \to \infty$, since p is constant in (0,1), it implies that $np \to \infty, n(1-p) \to \infty$ and x should be at the vicinity of n

Thus, let $\delta = x - np$, then $x = \delta + np$

$$ln[f(x)] = ln\left[\frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right]\right)}{\sqrt{2\pi x}\left(\frac{x}{e}\right)^x\left(\left[1+\mathcal{O}\left(\frac{1}{x}\right)\right]\right)\sqrt{2\pi(n-x)}\left(\frac{n-x}{e}\right)^{n-x}\left(\left[1+\mathcal{O}\left(\frac{1}{n-x}\right)\right]\right)}p^x(1\right.$$

$$\left.-p)^{n-x}\right]$$

$$= ln\left[\frac{\sqrt{\frac{n}{2\pi x(n-x)}}n^n}{x^x(n-x)^{n-x}}p^x(1-p)^{n-x}\right] + ln\left(\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right]\right) - ln\left(\left[1+\mathcal{O}\left(\frac{1}{x}\right)\right]\right)$$

$$- ln\left(\left[1+\mathcal{O}\left(\frac{1}{n-x}\right)\right]\right)$$

$$= ln\left(\sqrt{\frac{n}{2\pi x(n-x)}}n^n\left(\frac{p}{x}\right)^x\left(\frac{1-p}{n-x}\right)^{n-x}\right)$$

$$= ln\left(\sqrt{\frac{n}{2\pi(\delta+np)\big(n-(\delta+np)\big)}}n^x\left(\frac{p}{\delta+np}\right)^x n^{n-x}\left(\frac{1-p}{n-(\delta+np)}\right)^{n-x}\right)$$

$$= ln\left(\sqrt{\frac{n}{2\pi(\delta+np)\big(n-(\delta+np)\big)}}\left(\frac{np}{\delta+np}\right)^x\left(\frac{n(1-p)}{n(1-p)-\delta}\right)^{n-x}\right)$$

$$= ln\left(\sqrt{\frac{n}{2\pi(\delta+np)\big(n-(\delta+np)\big)}}\left(\frac{\delta+np}{np}\right)^{-x}\left(\frac{n(1-p)-\delta}{n(1-p)}\right)^{x-n}\right)$$

$$= ln\left(\sqrt{\frac{n}{2\pi(\delta+np)(n(1-p)-\delta)}}\right) - (\delta+np)ln\left(1+\frac{\delta}{np}\right)$$

$$- (n(1-p)-\delta)ln\left(1-\frac{\delta}{n(1-p)}\right)$$

By Taylor expansion:

$$\ln(1+n) = n - \frac{n^2}{2} + \mathcal{O}(n^3)$$

$$ln[f(x)] = ln\left(\sqrt{\frac{n}{2\pi(\delta+np)(n(1-p)-\delta)}}\right) - (\delta+np)\left[\left(\frac{\delta}{np}\right) - \frac{\left(\frac{\delta}{np}\right)^2}{2} + \mathcal{O}\left(\left(\frac{\delta}{np}\right)^3\right)\right]$$

$$- (n(1-p)-\delta)\left(-\frac{\delta}{n(1-p)} - \frac{\left(\frac{\delta}{n(1-p)}\right)^2}{2} + \mathcal{O}\left(\left(\frac{\delta}{n(1-p)}\right)^3\right)\right)$$

$$As\ n \to\ \infty,$$

$$ln\left(\sqrt{\frac{n}{2\pi(\delta+np)(n(1-p)-\delta)}}\right) = ln\left(\sqrt{\frac{1}{2\pi p(1-p)}}\right)$$

$$As\ n \to\ \infty,$$

$$(\delta+np)\left[\left(\frac{\delta}{np}\right) - \frac{\left(\frac{\delta}{np}\right)^2}{2} + \mathcal{O}\left(\left(\frac{\delta}{n}\right)^3\right)\right]$$

$$= \delta\left[\left(\frac{\delta}{np}\right) - \frac{\left(\frac{\delta}{np}\right)^2}{2} + \mathcal{O}\left(\left(\frac{\delta}{n}\right)^3\right)\right] + np\left[\left(\frac{\delta}{np}\right) - \frac{\left(\frac{\delta}{np}\right)^2}{2} + \mathcal{O}\left(\left(\frac{\delta}{n}\right)^3\right)\right]$$

$$= 0 + \delta - \frac{\delta^2}{2np} + np\mathcal{O}\left(\left(\frac{\delta}{n}\right)^3\right) = \delta - \frac{\delta^2}{2np} + \mathcal{O}\left(\frac{\delta^3}{n^2}\right)$$

$$As\ n \to\ \infty,$$

$$(n(1-p)-\delta)\left(-\frac{\delta}{n(1-p)}-\frac{\left(\frac{\delta}{n(1-p)}\right)^2}{2}+\mathcal{O}\left(\left(\frac{\delta}{n(1-p)}\right)^3\right)\right)$$

$$=-\delta\left(-\frac{\delta}{n(1-p)}-\frac{\left(\frac{\delta}{n(1-p)}\right)^2}{2}+\mathcal{O}\left(\left(\frac{\delta}{n(1-p)}\right)^3\right)\right)$$

$$+n(1-p)\left(-\frac{\delta}{n(1-p)}-\frac{\left(\frac{\delta}{n(1-p)}\right)^2}{2}+\mathcal{O}\left(\left(\frac{\delta}{n(1-p)}\right)^3\right)\right)$$

$$=0-\delta-\frac{\delta^2}{2n(1-p)}+\mathcal{O}\left(\frac{\delta^3}{n^2}\right)$$

$$ln[f(x)]=ln\left(\sqrt{\frac{1}{2\pi p(1-p)}}\right)+\left(\delta-\frac{\delta^2}{2np}+\mathcal{O}\left(\frac{\delta^3}{n^2}\right)\right)+\left(-\delta-\frac{\delta^2}{2n(1-p)}+\mathcal{O}\left(\frac{\delta^3}{n^2}\right)\right)$$

$$=ln\left(\sqrt{\frac{1}{2\pi p(1-p)}}\right)-\frac{\delta^2}{2np}-\frac{\delta^2}{2n(1-p)}+\mathcal{O}\left(\frac{\delta^3}{n^2}\right)$$

$$=ln\left(\sqrt{\frac{1}{2\pi p(1-p)}}\right)-\frac{\delta^2}{2np(1-p)}+\mathcal{O}\left(\frac{\delta^3}{n^2}\right)$$

$$As\ n \to\ \infty,$$

$$f(x) = \sqrt{\frac{1}{2\pi p(1-p)}} exp\left(-\frac{(x-np)^2}{2np(1-p)}\right)$$

Since $\mu = np$ and $\sigma^2 = np(1-p)$ for binomial distribution, the on-chromatography analyte concentration distribution approaches Gaussian distribution as theoretical plate number approaches infinity.

4.3.2 Gaussian Approximation of Outflow Analyte Concentration Distribution

In chapter 2, we have proved that the outflow chromatography analyte concentration distribution has the same formula as negative binomial distribution. The common knowledge from empirical curve fitting is that outflow chromatography analyte concentration distributions have the same formula as the Gaussian distribution as theoretical plate number approaches infinity [40]. We assume that this observation is true and let us prove it.

Assumptions:

Since partition coefficient ρ is assumed to be not close to 0 or infinity, we have the proportion constant $\lambda = \frac{\rho}{1+\rho}$ not close to 0 or 1. And the theoretical plate number r approaches infinity.

The outflow quantity of analyte as function of stage is the mobile phase of the last plate (the plate at the cutoff) which is:

$$M_i^j = \binom{i-1}{j-1} \lambda^j (1-\lambda)^{i-j}$$

For simplicity, we switch to commonly known negative binomial distribution conventions for parameters for its mathematical formula:

Let $f(y)$has the same formula as $NB(r,p)$, and assume that $r \to \infty$

$$f(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r}, y \in \{r, r+1, \dots\}, r \in \{1,2, \dots\}$$

Let us prove that $Y \sim N\left(\frac{r}{p}, \frac{r(1-p)}{p^2}\right)$, as $r \to \infty$

Note: as $r \to \infty$ we have $rp \to \infty, r(1-p) \to \infty$

$$Let\ X_i \sim iid\ Geo(p), then\ Y = \sum_{i=1}^{r} X_i$$

$$f(x) = p(1-p)^{x-1}, x\{1,2, \dots\}\ with\ mean\ \mu_X = \frac{1}{p}$$

Moment generation function is:

$$m(t) = E\left(e^{t(X-\mu_X)}\right)$$

$$m(0) = E\left(e^{0(X-\mu_X)}\right) = 1$$

By taylor expansion: $e^{t(X-\mu_X)} = 1 + t(X_i - \mu_X) + \frac{\left(t(X_i-\mu_X)\right)^2}{2!} + \cdots$

$$m'(0) = \left.\frac{\partial m(t)}{\partial t}\right|_{t=0} = \left.\frac{\partial E e^{t(X-\mu_X)}}{\partial t}\right|_{t=0} = \left.\frac{\partial E\left(1 + t(X_i - \mu_X) + \frac{\left(t(X_i-\mu_X)\right)^2}{2!} + \cdots\right)}{\partial t}\right|_{t=0}$$

$$= E\left(\left.\frac{\partial\left(1 + t(X_i - \mu_X) + \frac{\left(t(X_i-\mu_X)\right)^2}{2!} + \cdots\right)}{\partial t}\right|_{t=0}\right) = E(X_i - \mu_X) = 0$$

$$m''(0) = \frac{\partial^2 m(t)}{\partial t^2}\bigg|_{t=0} = \frac{\partial^2 E\left(1 + t(X_i - \mu_X) + \frac{\left(t(X_i - \mu_X)\right)^2}{2!} + \frac{\left(t(X_i - \mu_X)\right)^3}{3!} + \cdots\right)}{\partial t^2}\bigg|_{t=0}$$

$$= E\left[\frac{\partial^2\left(1 + t(X_i - \mu_X) + \frac{\left(t(X_i - \mu_X)\right)^2}{2!} + \frac{\left(t(X_i - \mu_X)\right)^3}{3!} + \cdots\right)}{\partial t^2}\bigg|_{t=0}\right]$$

$$= E\left(\frac{\partial^2\left(t(X_i - \mu_X)\right)^2}{2\partial t^2}\right) = E\left(\frac{2}{2}((X_i - \mu_X)^2)\right) = E(X_i - \mu_X)^2 = Var(X_i) = \frac{(1-p)}{p^2}$$

By Taylor expansion:

$$m(t) = 1 + m'(0)t + \frac{m''(0)}{2!}t^2 + \frac{m'''(0)}{3!}t^3 + \cdots$$

By Taylor theorem, there exist a $\zeta(0 < \zeta < t)$ such that:

$$\frac{m''(\zeta)}{2!}t^2 = \frac{m''(0)}{2!}t^2 + \frac{m^{(3)}(0)}{3!}t^3 + \cdots$$

Thus we have:

$$m(t) = 1 + m'(0)t + \frac{m''(\zeta)}{2!}t^2 = 1 + 0 + \frac{m''(0)}{2!}t^2 + \frac{m''(\zeta)}{2!}t^2 - \frac{m''(0)}{2!}t^2$$

$$= 1 + 0 + \frac{\sigma_X^2}{2}t^2 + \frac{m''(\zeta) - \sigma_X^2}{2}t^2$$

$$Let\ Z_r = \frac{\sum_{i=1}^r (X_i - \mu_X)}{\sigma_X\sqrt{r}}\ then\ M_{Z_r}(t) = M_{\frac{\sum_1^r (X_i - \mu_X)}{\sigma_X\sqrt{r}}}(t) = Ee^{\frac{\sum_1^r (X_i - \mu_X)}{\sigma_X\sqrt{r}}t} = \prod_{i=1}^{r} Ee^{\frac{(X_i - \mu_X)}{\sigma_X\sqrt{r}}t}$$

$$= \prod_{i=1}^{r} M_{X_i-\mu_X}\left(\frac{t}{\sigma_X\sqrt{r}}\right) = \left(M_{X_i-\mu_X}\left(\frac{t}{\sigma_X\sqrt{r}}\right)\right)^r = \left(m\left(\frac{t}{\sigma_X\sqrt{r}}\right)\right)^r$$

$$= \left(1 + \frac{\sigma_X^2\left(\frac{t}{\sigma_X\sqrt{r}}\right)^2}{2} + \frac{m''(\zeta) - \sigma_X^2}{2}\left(\frac{t}{\sigma_X\sqrt{r}}\right)^2\right)^r \; where\; \zeta \;\; (0 < \zeta < t)$$

$\lim_{r\to\infty} \frac{t}{\sigma_X\sqrt{r}} \to 0$ thus by sandwich theorem: $\zeta \to 0\; as\; r \to \infty$

$$\lim_{r\to\infty}(m''(\zeta) - \sigma_X^2) = \lim_{\zeta\to 0}(m''(\zeta) - \sigma_X^2) = m''(0) - \sigma_X^2 = 0$$

$$\lim_{r\to\infty} M_{Z_r}(t) = \lim_{r\to\infty}\left(1 + \frac{\sigma^2\left(\frac{t}{\sigma\sqrt{r}}\right)^2}{2}\right)^r = \lim_{r\to\infty}\left(1 + \frac{t^2}{2n}\right)^r = e^{\frac{t^2}{2}}$$

Therefore, $\lim_{r\to\infty} Z_r \equiv Z{\sim}N(0,1)$ by matching the moment generating function .

$Y = \sum_{i=1}^{r} X_i = \sqrt{r}\sigma_X Z_r + r\mu_X$

$r\mu_X = \frac{r}{p}, r\sigma_X^2 = r\frac{(1-p)}{p^2}$

Therefore Y has same formula as the distribution $N\left(\frac{r}{p}, r\frac{(1-p)}{p^2}\right)$

As the number theoretical plates approaches infinity, we can approximate the outflow chromatography analyte concentration distribution as following:

$f(y) = \frac{p}{\sqrt{2\pi r(1-p)}}\exp\left(-\frac{p\left(y-\frac{r}{p}\right)^2}{2r(1-p)}\right), y \in (r,\infty), as\; r \to \infty$

4.4. Parameter Estimation

4.4.1 Method of Moments Estimator (MOM)

The method of moments (MOM) estimator can be obtained by solving the following simultaneous equations to (match the first and second moments) observed to their theoretical values, or:

$$\begin{cases} m_1 = \dfrac{\sum_i x_i}{n} \;\; match\, with \;\; \mu_1' = \dfrac{r}{p} \\ m_2 = \dfrac{\sum_i x_i^2}{n} \;\; match\, with \;\; \mu_2' = \left(\dfrac{r}{p}\right)^2 + \dfrac{r(1-p)}{p^2} \end{cases}$$

$$\begin{cases} \bar{x} = \dfrac{r}{p} \\ \overline{x^2} = \left(\dfrac{r}{p}\right)^2 + \dfrac{r(1-p)}{p^2} \end{cases}$$

$$\begin{cases} r = \bar{x}p \\ \overline{x^2} = \left(\dfrac{r}{p}\right)^2 + \dfrac{r(1-p)}{p^2} \end{cases}$$

So,

$$\overline{x^2} = \left(\frac{\bar{x}p}{p}\right)^2 + \frac{\bar{x}p(1-p)}{p^2}$$

The solution is:

$$\begin{cases} \tilde{r} = \dfrac{(\bar{x})^2}{\overline{x^2} + \bar{x} - (\bar{x})^2} \\ \tilde{p} = \dfrac{\bar{x}}{\overline{x^2} + \bar{x} - (\bar{x})^2} \end{cases}$$

From the transformed data, the sample mean and sample square mean is

$$\bar{x} = 16.72 \; and \; \overline{x^2} = 485.06$$

Thus the MOM estimator for the total plate number and proportional constant is

$$\begin{cases} \tilde{r} = \dfrac{(\bar{x})^2}{\overline{x^2} + \bar{x} - (\bar{x})^2} = 53 \\ \tilde{p} = \dfrac{\bar{x}}{\overline{x^2} + \bar{x} - (\bar{x})^2} = 0.875 \end{cases}$$

Proof of the invariance property of MOM in weight-frequency conversion:

Assume that instead of 1 milligram per count, we use w milligram per count as conversion criteria. Then the total number of each x_i in the dataset become the number of x_i in original converted dataset multiply w, therefore $(\sum_i x_i)_{new} = w(\sum_i x_i)_0$ and $(\sum_i x_i^2)_{new} = w(\sum_i x_i^2)_0$ and $n_{new} = wn_0$

$$\begin{cases} m_{1_{new}} = \dfrac{(\sum_i x_i)_{new}}{n_{new}} = \dfrac{w(\sum_i x_i)_0}{wn_0} = \dfrac{(\sum_i x_i)_0}{n_0} = m_{10} \\ m_{2_{new}} = \dfrac{(\sum_i x_i^2)_{new}}{n_{new}} = \dfrac{(\sum_i x_i^2)_0}{n_0} = m_{20} \end{cases}$$

Thus invariance property of MOM is proved for the analyte weight to frequency conversion.

4.4.2 Maximum Likelihood Estimator (MLE)

The maximum likelihood estimator can be solved by numerical method such as Newton-Ralphson

Denote $\boldsymbol{\theta} = (r, p)^T$, then the likelihood function is :

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{m} f(y_i|r,p) = \prod_{i=1}^{m} \binom{y_i - 1}{r-1} p^r (1-p)^{y-r} = \prod_{i=1}^{m} \frac{\Gamma(y_i)}{\Gamma(r)\Gamma(y_i - r + 1)} p^r (1-p)^{y_i - r}$$

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = lnL(\boldsymbol{\theta}|\boldsymbol{y})$$

$$= \sum_{i=1}^{m} ln\Gamma(y_i) + \sum_{i=1}^{m} (y_i - r)\ln(1-p) + mrln(p) - mln\Gamma(r)$$

$$- \sum_{i=1}^{m} ln\Gamma(y_i - r + 1)$$

$$\frac{\partial l}{\partial r} = -mln(1-p) + mln(p) - mDigamma(r) + \sum_{i=1}^{m} Digamma\left(y_i - r + 1\right)$$

$$\frac{\partial l}{\partial p} = -\frac{\sum_{i=1}^{m}\left(y_i - r\right)}{1-p} + \frac{mr}{p}$$

$$\frac{\partial^2 l}{\partial r^2} = \frac{\partial\left(\frac{\partial l}{\partial r}\right)}{\partial r} = -mTrigamma(r) - \sum_{i=1}^{m} Trigamma\left(y_i - r + 1\right)$$

$$\frac{\partial^2 l}{\partial p^2} = \frac{\partial\left(\frac{\partial l}{\partial p}\right)}{\partial p} = -\frac{\sum_{i=1}^{m}\left(y_i - r\right)}{(1-p)^2} - \frac{mr}{p^2}$$

$$\frac{\partial^2 l}{\partial r \partial p} = \frac{\partial\left(\frac{\partial l}{\partial p}\right)}{\partial r} = \frac{m}{p} - \frac{m}{1-p}$$

The gradient vector is

$$\boldsymbol{g} = \frac{\partial l}{\partial \boldsymbol{\theta}^T} = \left(\frac{\partial l}{\partial r}, \frac{\partial l}{\partial p}\right)^T = \left(\frac{\partial l}{\partial r}, -\frac{\sum_{i=1}^{m}\left(y_i - r\right)}{1-p} + \frac{mr}{p}\right)^T$$

and the hessian matrix is

$$\boldsymbol{H} = \frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{pmatrix} \frac{\partial^2 l}{\partial r^2} & \frac{\partial^2 l}{\partial r \partial p} \\ \frac{\partial^2 l}{\partial r \partial p} & \frac{\partial^2 l}{\partial p^2} \end{pmatrix}$$

$$= \begin{pmatrix} -mTri\Gamma(r) - \sum_{i=1}^{m} Tri\Gamma(y_i - r + 1) & \frac{m}{p} - \frac{m}{1-p} \\ \frac{m}{p} - \frac{m}{1-p} & \frac{\sum_{i=1}^{m}(y_i - r)}{(1-p)^2} - \frac{mr}{p^2} \end{pmatrix}$$

We can set initial guess of maximum likelihood estimator by using MOM estimator

$$\widehat{\boldsymbol{\theta}}_{(0)} = \widetilde{\boldsymbol{\theta}} = \begin{pmatrix} \tilde{r} \\ \tilde{p} \end{pmatrix} = \begin{pmatrix} 53 \\ 0.875 \end{pmatrix}$$

The recursion relationship is:

$$\widehat{\boldsymbol{\theta}}_{(\boldsymbol{t+1})} = \widehat{\boldsymbol{\theta}}_{(\boldsymbol{t})} - \boldsymbol{H}^{-1}\boldsymbol{g} = \widehat{\boldsymbol{\theta}}_{(\boldsymbol{t})} - \begin{pmatrix} \frac{\partial^2 l}{\partial r^2} & \frac{\partial^2 l}{\partial r \partial p} \\ \frac{\partial^2 l}{\partial r \partial p} & \frac{\partial^2 l}{\partial p^2} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial l}{\partial r} \\ \frac{\partial l}{\partial p} \end{pmatrix}$$

And the convergence criteria:

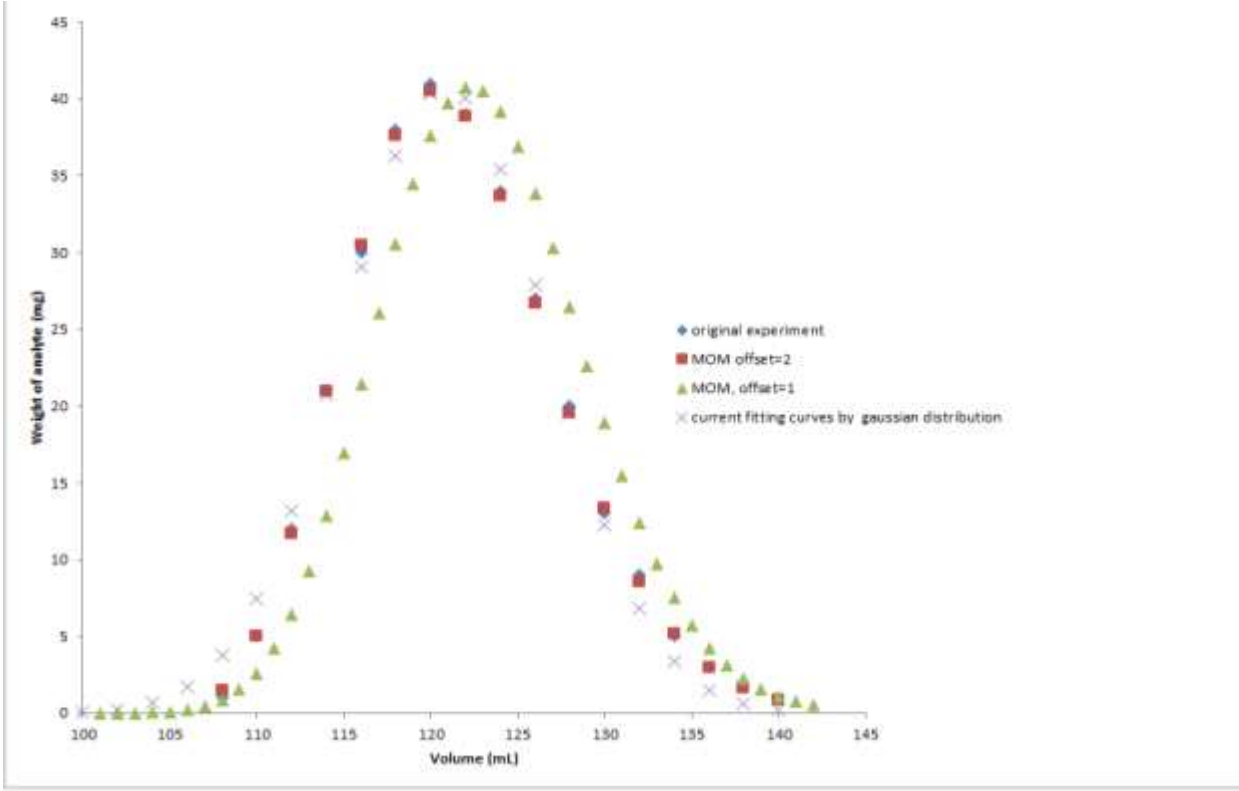
$$\left| \frac{\left((lnL)_{(t+1)} - (lnL)_{(t)} \right)}{(lnL)_{(t)}} \right| \leq \delta$$

4.5. Simulation

Simulation of the outflow peaks from chromatography was conducted by negative binomial model using the parameters estimated by MOM in previous section with correct offset of 2mL/plate. An unadjusted offset 1mL/plate was also used, and this served to demonstrate the

importance of the correct offset specification. These results were compared to original experimental data and also to results derived under the current standard approach of empirically matching to a Gaussian model [12]. Results are presented in Figure 4.1. All analyte concentration distributions are overlayed in same plot.

Figure 4.1Chromatography analyte concentration distribution of compound 1, original experimental data compared to Gaussian model and model with same formula as negative binomial distribution.



The simulation of the outflow peaks modeled utilizing negative binomial distribution with parameter obtained by MOM estimator using correct offset renders the closest results to the actual, original data. The simulation of outflow analyte concentration distribution with Gaussian model is slightly off in that it did not catch the skewness characteristic of the experimental outflow analyte concentration distribution. The simulation of the outflow analyte concentration

distributions modeled utilizing negative binomial distribution with the unadjusted offset is largely deviated from the experimental outflow data, particularly as compared to the negative binomial simulation with correct offset currently and also compared to the prevailing Gaussian model. Therefore it is important to find offset correctly so that chromatography analyte concentration distribution parameters can estimated more accurately.

Objective measures to compare the simulated analyte concentration distributions is examined by using a "chromatogram information criterion" (CIC), which is the sum of squares of deviation of expected frequencies from the observed frequencies normalized by the expected frequency. It is the same formula as the Pearson χ^2 goodness-of-fit test statistics; however, we do not compare this to χ^2 distribution as done in context of goodness-of-fit testing because the sample size of data created from the chromatogram can be arbitrarily increased multiplicatively.

We use this formula with fixed total frequency of data points produced from chromatogram to compare different models that are estimated. The larger number of this criterion indicate a worse fit of the model that generate these data and it is a relative comparison between model and is not absolute comparison to χ^2 distribution.

$$CIC = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

Here, the f_i indicates the observed frequencies in transformed data, and e_i indicates the expected frequencies. If we assume that the correct model is the negative binomial, then expected frequencies is shown in table 4.4

Table 4.4 Expected frequencies estimated by negative binomial model with correct offset and parameters estimated by MOM and compared to observed frequencies

	f_i	e_i		f_i	e_i
54	1	1.452	63	27	26.709
55	5	5.004	64	20	19.569
56	12	11.709	65	13	13.356
57	21	20.928	66	9	8.547
58	30	30.471	67	5	5.16
59	38	37.632	68	3	2.952
60	41	40.533	69	2	1.608
61	39	38.862	70	1	0.834
62	34	33.681			

$$CIC_{MOM\_adjusted} = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = 0.3846$$

The $CIC_{\ MOM\_adjusted}$ of 0.3846 is very small, which indicate that probably of outflow peak the deviate from negative binomial model with correct offset and parameters estimated by MOM is very small.

Similarly we can calculate CIC criterion for the Gaussian model and the negative binomial model without adjustment by offset and parameters estimated by MOM.

$$CIC_{Gaussian} = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = 12.78$$

$$CIC_{MOM\_unadjusted} = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = 25.64$$

Based on CIC criterion, the formula for the negative binomial model with correct offset and parameters estimated by MOM best approximated the chromatography outflow peak data.

4.6. Conclusion

In this work, we have successfully estimated both parameters that determine the shape and location of chromatography peak (partition coefficient and total number of theoretical plate) simultaneously by using statistical method without any additional experiment. The comparison the simulated outflow peaks using current prevailing Gaussian formula, unadjusted negative binomial formula, to the negative binomial formula using parameters estimated by MOM with correct offset shows that the negative binomial formula using parameters estimated by MOM with correct offset most closely matched to the experimental data. We have also proved that as total plate number approach infinity, and the proportion constant not approaching 0 or 1 the outflow distribution (negative binomial distribution) converges to Gaussian distribution.

In the future work, the maximum likelihood estimator (MLE) for the chromatography parameters will be computed by Newton-Ralphson method described in chapter 4 using SAS proc IML. The simulated peak based on MLE will be compared with that based on MOM estimator and original data, as well as the commonly used Gaussian formula. The separation process of the chromatography that separates of several analytes will be modeled similarly. The chromatography data would contain more than one analyte concentration distribution. This type of chromatography data containing multiple components will be analyzed in a similar way and the partition coefficients of multiple components will be estimated by both MOM and ML method.

Page left intentionally blank

Chapter 5: Summary

In summary, this dissertation work not only developed the mathematical model for chromatography separation process and applied statistical method in estimation of chromatography parameters, but also proposed and proved the relation (binomial-negative binomial theorem) between on-chromatography and outflow analyte concentration distributions. Furthermore, this dissertation work proved that for large theoretical plate number, the formula for the Gaussian distribution provides good approximations of both on-chromatography analyte concentration distributions and outflow analyte concentration distributions.

Chapter 2 proposed and proved that for chromatography with relatively small number of theoretical plate, the on-chromatography analyte concentration distribution mathematically matches the binomial distribution, and outflow analyte concentration distribution matches the formula for the negative binomial distribution. In this chapter, the chromatography table was utilized for visualization of chromatography process and the on-chromatography and outflow analyte concentration distributions. Simulations conducted based on the principle of chromatography shows that the proposed negative binomial distribution formula is more suited to fit the outflow analyte concentration distribution than the prevailing Gaussian distribution formula.

Chapter 3 proposed and proved binomial-negative binomial theorem to elucidate the relationship between the discrete on-chromatography and outflow analyte concentration distributions. In this chapter, the on-chromatography analyte concentration distribution and outflow analyte concentration distribution were plotted in a 3D graph to facilitate the

visualization the corresponding chromatography process and understanding relationship between the discrete on-chromatography and outflow analyte concentration distributions.

Chapter 4 is combination of theoretical development in asymptotic chromatography analyte concentration distribution and the application of statistical method for chromatography parameter estimation. In this chapter, the developed asymptotic chromatography analyte concentration distribution theory build a bridge between discrete mathematical model for the chromatography with relatively smaller number of theoretical plates and the continuous Gaussian distribution model that majority researchers are using to approximate the chromatography analyte concentration distributions. The Gaussian distribution formula is only valid for approximate the outflow peaks when the number of theoretical plate is sufficiently large. We have also established the method for transformation of chromatography data to statistical sampling data so that statistical method of parameter estimation can be conducted. The simulation shows that the negative binomial distribution's mathematical formula using parameters estimated by MOM using correct offset best approximate the outflow analyte concentration distribution from actual experimental data in comparison to: the negative binomial distribution's formula using parameters estimated by MOM without correct offset; and prevailing Gaussian distribution formula matching method.

Reference

[1] Tswett, M. *Berichte der Deutschen botanischen Gesellschaft*, (1906) 24 316

[2] Still, W. C.; Kahn, M.; Mitra, A. J. Org. Chem. (1978) 43 (14): 2923–2925.

[3] Han, Zhenwei; He, Zhimin; Yu, Guocong Sepu (1997), 15(6), 532-533.

[4] Eriksen, Stuart P. American Journal of Pharmaceutical Education (1965), 29(4), 518-33.

[5] Yang, Z.-H.; Wnag, B.; Liang Y.-Z.; Xie, G.-X.; Ren, X-M. J. Sep. Sci. 2011, 34, 978-986

[6] Coolidge, J. L.; The American Mathematical Monthly, (1949), 56, (3) 147-157

[7] Broadhurst, Hugh A.; Rein, Peter W. Zuckerindustrie (2003), 128(2), 96

[8] Caninde de Sousa, F.; Padilha, Carlos E. A.; Chiberio,A.S.; Ribeiro, V. T.; Martins, D. A.; Araujo de Oliveira, J.; Ribeiro de Macedo, G. and Silvino dos Santos, E. Separation and Purification Technology (2016), 164, 34

[9] Still, W. C.; Kahn, M.; Mitra, A. *J. Org. Chem.* (1978) 43 (14): 2923–2925.

[10] Han, Zhenwei; He, Zhimin; Yu, Guocong *Sepu* (1997), 15(6), 532-533

[11] (a) Grimalt, Joan; Iturriaga, Hortensia; Olive, Joaquim *Analytica Chimica Acta* (1987), 201, 193-205.

(b) Eriksen, Stuart P. *American Journal of Pharmaceutical Education* (1965), 29(4), 518-33.

(c) Xiu, Guo-Hua; Li, Ping *Chemical Engineering Science* (1999), 54(3), 377-387.

(d) Kalambet, Yuri; Kozmin, Yuri; Mikhailova, Ksenia; Nagaev, Igor; Tikhonov, *Pavel Journal of Chemometrics* (2011), 25(7), 352-356.

(e) Pai, Su-Cheng *Journal of Chromatography A* (2004), 1028(1), 89-103

(f) Broadhurst, Hugh A.; Rein, Peter W. *Zuckerindustrie* (2003), 128(2), 96-99.

[12] (a) Foley, Joe P.*Analytical Chemistry* (1987), 59(15), 1984-7

(b) Buys, T. S.; De Clerk, K. *Separation Science* (1972), 7(4), 441-8.

[13] Liu, Shu-Jiang; Chen, Zhan-Ying; Chang, Yin-Zhong; Wang, Shi-Lian; Qi, Li; Fan, Yuan-Qing *Journal of Chromatography A* (2013), 1311, 183-187

[14] (a) Kanev, A. S.; Rysev, O. A.; Chechevichkin, V. N. *Zavodskaya Laboratoriya* (1981), 47(4), 13-15.

(b) Belenkii, B. G.; Nesterov, V. V.; Smirnov, M. M. *Zhurnal Fizicheskoi Khimii* (1968), 42(6), 1484-9

[15] (a) Leboda, Roman *Journal of Chromatography* (1979), 178(2), 369-86.

(b) Rudenko, B. A.; Metlyaeva, S. Ya.; Il'kova, E. L. *Zhurnal Analiticheskoi Khimii* (1970), 25(4), 670-8

[16] (a) Grimalt, Joan; Iturriaga, Hortensia; Olive, Joaquim *Analytica Chimica Acta* (1987), 201, 193-205.

(b) Liu, Shu-Jiang; Chen, Zhan-Ying; Chang, Yin-Zhong; Wang, Shi-Lian; Qi, Li; Fan, Yuan-Qing *Journal of Chromatography A* (2013), 1311, 183-187

[17] K. Kaczmarski *Computers & Chemical Engineering* 1996, 20 (1) 49–64

[18] Broadhurst, Hugh A.; Rein, Peter W. *Zuckerindustrie* (2003), 128(2), 96-99.

[19] Kalambet, Yuri; Kozmin, Yuri; Mikhailova, Ksenia; Nagaev, Igor; Tikhonov, Pavel *Journal of Chemometrics* (2011), 25(7), 352-356

[20] Yang, Z.-H.; Wnag, B.; Liang Y.-Z.; Xie, G.-X.; Ren, X-M. *J. Sep. Sci.* 2011, 34, 978-986

[21] (a) Coolidge, J. L.; *The American Mathematical Monthly*, (1949), 56, (3) 147-157

(b) Ross, G. J.; Preece, D. A. *Journal of the Royal Statistical Society. Series D* (1985), 34, (3) 323-335

(c) Cacoullos, T; Papageorgiou, H.; *Journal of Applied Probability*, (1982), 19, (3) 742-743

[22] Allison, P. D.; Waterman, R. P. *Sociological Methodology*, (2002), 32, 247-265

[23] Han, Z.; He, Z.; Yu, G. *Chinese Journal of Chromatography (Sepu)* (1997), 15(6), 532-533.

[24] Yang, Z.-H.; Wnag, B.; Liang Y.-Z.; Xie, G.-X.; Ren, X-M. *J. Sep. Sci.* 2011, 34, 978-986

[25] Kalambet, Y.; Kozmin, Y.; Mikhailova, K.; Nagaev, I.; Tikhonov, P.; *Journal of Chemometrics* (2011), 25(7), 352-356.

[26] Denizot, F. C.; Delaage, M. A. *Proceedings of the National Academy of Sciences of the United States of America* (1975), 72(12), 4840-3.

[27] Casella, G.; Berger, R. L. *Statistical Inference* (2002) Thomson Press

[28] Pai, Su-Cheng *Journal of Chromatography A* (2004), 1028, 89

[29] Netopilik, M. *Journal of Chromatography A* (2006), 1133, 95

[30] Grimalt, Joan; Iturriaga, Hortensia; Olive, Joaquim *Analytica Chimica Acta* (1987), 201, 193

[31] Eriksen, Stuart P. *American Journal of Pharmaceutical Education* (1965), 29(4), 518

[32] Xiu, Guo-Hua; Li, Ping *Chemical Engineering Science* (1999), 54(3), 377

[33] Yang, Z.-H.; Wnag, B.; Liang Y.-Z.; Xie, G.-X.; Ren, X-M. *J. Sep. Sci.* (2011), 34, 978

[34] Broadhurst, Hugh A.; Rein, Peter W. *Zuckerindustrie* (2003), 128(2), 96

[35] Chin,Y.; Weber, W. and Voic, T. C. *War. Res.* (1986) 20 (11), 1433

[36] Caninde de Sousa, F.; Padilha, Carlos E. A.; Chiberio,A.S.; Ribeiro, V. T.; Martins, D. A.; Araujo de Oliveira, J.; Ribeiro de Macedo, G. and Silvino dos Santos, E. *Separation and Purification Technology* (2016), 164, 34

[37] Liu, S.; Chen, Z.; Chang,Y.; Wang, S.; Qi, L.; Fan, Y. *Journal of Chromatography A* (2013), 1311, 183

[38] Bai, X.; Chen, X.; Dias, J. R. and Sandreczki, T.C *Tetrahedron Letters*, (2013), 54, 1711

[39] Liu, Shu-Jiang; Chen, Zhan-Ying; Chang, Yin-Zhong; Wang, Shi-Lian; Qi, Li; Fan, Yuan-Qing *Journal of Chromatography A* (2013), 1311, 183

[40] Dinov, I. D.; Christou, N.; and Sanchez, J. *Journal of Statistics Education* (2008),16, 34

Appendix

Chapter 2 Codes

SAS code for chromatography analyte concentration distribution simulation

```
/*************************************************************/
/*********** chromatography peak simulation ****************/
/***********          10/09/2015                 *************/
/***********         by xueyi chen               *************/
/*************************************************************/

Proc iml;
/*specify parameters*/
p=0.8754;
n=500;
r=53.2;
m=0*J(n,n);
s=0*J(n,n);
t=0*J(n,n);

m[1,1]=p;
S[1,1]=1-p;
t[1,1]=1;
do i =2 to n;
 do j=1 to i;

  if j=1 then do;
   m[i,j]=p*s[i-1,j];
   s[i,j]=(1-p)*s[i-1,j];

  end;

  else do;
   m[i,j]=p*(m[i-1,j-1]+s[i-1,j]);
   s[i,j]=(1-p)*(m[i-1,j-1]+s[i-1,j]);
  end;
  t[i,j]=m[i,j]+s[i,j];

 end;
end;



outflow=m[,r];
*print outflow;
t=n-r+1;
sigma_sq=0.001;
outsim=0*j(t,1);
do i =1 to t;
```

```
err=rand('NORMAL',0,sigma_sq);
if err<-outflow[i+r-1] then do;
err1=0;
end;
else do;
err1=err;
end;

outsim[i]=outflow[i+r-1]+err1;
end;

*print outsim;
out_theo=0*j(t,1);

do i=1 to t;
out_theo[i]=PDF('NEGBINOMIAL',i-1,p,r);
end;

stage=0*j(t,1);
do i=1 to t;
stage[i]=i+r;
end;
ID1=1*j(n+1,1);
ID2=2*j(n+1,1);

theo_dat=(stage||out_theo);
sim_dat=(stage||out_sim);

print dat;

dv_stage=0*j(r,1);
do i=1 to r;
dv_stage[i]=i;
end;
dv_theo=0*j(r,1);
dv_sim=0*j(r,1);

dat1=(((dv_stage||dv_theo)//(stage||out_theo))||ID1)//(((dv_stage||dv_sim)//(
stage||outsim))||ID2);
*print dat1;




stage1=0*j(n+1,1);

do i=1 to n+1;
stage1[i]=i;
end;

mode=(1-p)*(r-1)/p;
Var=(1-p)*r/(p**2);
sd=sqrt(Var);

out_normal=0*j(n+1,1);
do i=1 to n+1;
```

```
a=i+t-(n+1);
out_normal[i]=PDF('NORMAL',a,mode,sd);
end;
ID3=3*j(n+1,1);

dat2=stage1||out_normal||ID3;
dat=dat1//dat2;
print dat;


create ChrSim from dat[colname={"stage" "outflow_distn" "ID"}];
append from dat;
close ChrSim;
print dat;
quit;

data ChrSim1;
set ChrSim;
if ID=1 then ID1='outflow_distn_by_theory';
if ID=2 then ID1='Negative_binomail_model';
if ID=3 then ID1='Gaussian_model';
drop ID;
run;


proc sgplot data=ChrSim1;
  scatter x=stage y=outflow_distn / group=ID1;
  YAXIS LABEL = 'peak intensity' GRID VALUES = (0 TO 0.5 BY 0.02);
  XAXIS LABEL = 'retention' GRID VALUES = (95 TO 180 BY 5);
run;

ods csv file='C:\passport_data\disertation\BNB_paper\comp1x.csv';
proc print data=ChrSim1 (firstobs=50 obs=72);
run;
ods csv close;


data ChrSim2;
set ChrSim1;
where stage>=50 and stage<=75;
run;


ods csv file='C:\passport_data\disertation\BNB_paper\comp1g.csv';
proc print data=ChrSim2 ;
run;
ods csv close;
```