Comparison of Several Percentile Definitions

White Paper

National Institute for Occupational Safety and Health (NIOSH)

December 2019

Daniel O. Stancescu
Division of Compensation Analysis and Support, NIOSH

A. Iulian Apostoaei, Brian A. Thomas
Oak Ridge Center for Risk Analysis, Inc. (ORRISK, Inc.)

Reviewed by Timothy D. Taulbee
Division of Compensation Analysis and Support, NIOSH

ABSTRACT

Claims filed under the Energy Employees Occupational Illness Compensation Program Act (EEOICPA) are evaluated using an estimated **p**robability of cancer **c**ausation (PC) for exposed workers. The PC is represented by a probability distribution function (pdf) from 2,000 or more samples produced using Monte-Carlo iterations to account for uncertainties in radiation dose and risk parameters representing the exposure situation, and the 99th percentile of this distribution is compared to the decision threshold of 50%.

There is no standard method to estimate a percentile (quantile) given multiple samples from a probability distribution, and different software packages use different methods. This report provides a survey of scientific literature on methods to calculate a sample percentile. Ten most used methods have been identified and tested for cases relevant to the EEOICPA compensation program. The performance of a percentile method differs, depending on the percentile level, sample size, type of underlying distribution, and magnitude of uncertainty and of positive skewness, but, in general, the larger the sample size, the smaller the difference between the values of the same percentile produced by different methods. Testing consisted of sampling from a single distribution or from the sum or product of two distributions selected from among the most common ones used in the NIOSH-IREP software. Sample percentiles obtained using the ten methods were compared with theoretical percentiles obtained analytically from the same selected distributions, and differences were evaluated using several statistical measures. For all tested scenarios, the sample percentile method used by default in the SAS software had the smallest relative bias and smallest relative root mean squared error at all the sample sizes, when estimating the 99th percentile. The sample percentile method used in NIOSH-IREP through the Analytica software tends to underestimate the theoretical 99th percentile in all the scenarios tested, especially for sample sizes lower than 20,000 values. For sample sizes greater than 20,000 the difference between methods becomes small enough and can be neglected. Thus, it is preferrable to use 20,000 or more samples in PC calculations using NIOSH-IREP for EEIOCPA claim evaluations.

TABLE OF CONTENTS

1.0 INTRODUCTION

The purpose of this document is to summarize and compare several definitions of sample percentiles (quantiles) and to clarify which of these definitions is used in the NIOSH-IREP software, for the purpose of computing the Probability of Causation (PC). The document provides a comparison of the percentile definitions used in the most common software packages and analyzes which of the percentile definitions produce the least biased results when sampling from one distribution, as well as in the process of combining two distributions through the Monte Carlo computations. Several simulations are performed to determine if any percentile definition has an advantage in terms of bias and variability at different sample sizes for the most common distributions used in the NIOSH-IREP software.

2.0 PERCENTILE DEFINITIONS

There is no standard method to estimate a percentile (quantile) from multiple samples obtained from a probability distribution. Multiple definitions are currently in use, with each software package using different methods to compute percentiles. All the probabilistic modeling and risk analysis software packages like Crystal Ball (Oracle 2010), @Risk (Palisade 2016), Analytica (Lumina 2015), and Model Risk (Vose 2012) have different methods implemented to compute percentiles. Statistical software packages like SAS or R have multiple definitions implemented to compute a percentile, with one method being used by default. There are several articles in statistical literature, dealing with the issue of which percentile definition to use. Hyndman and Fan (1996) evaluated nine different methods for computing the quantiles relative to six desirable properties and advised standardizing the quantile definition in statistical software packages. Langford (2006) compared a dozen different ways to compute quartiles used in different statistical packages. Makkonen and Pajari (2014) proposed a new method based on the use of true rank probabilities for defining sample quantiles. Kreutzmann (2018) analyzed estimating sample quantiles with the default definition in different software programs which can lead to very different results, due to the fact that software packages use different quantile definitions.

In the absence of a standard method to compute a sample percentile, there is a need to clarify how the 99th percentile is estimated in the NIOSH-IREP software, and how accurate this estimation is during the process of Monte Carlo computations used to determine Probability of Causation (PC) values.

Percentiles split a probability distribution into 100 intervals of equal probability, similar to deciles (which split data into 10 intervals), and quartiles (split the data into 4 intervals), and all of them are special cases of quantiles (which split the data into q intervals). Quantiles are defined by the inverse of the cumulative distribution function (cdf; $F(x)$), if the distribution is known and invertible. The quantile function Q is formally defined using the inverse function of the cdf

$$Q(p) = F^{-1}(p) = \inf\left\{x \in R : F(x) \geq p\right\}, \quad \text{for } 0 < p < 1,$$

where $F(x) = Prob(X \le x)$ is the cdf of the distribution X, and p is the quantile level, and inf represents the infimum of a set (in most cases the infimum is the smallest value in a set).

In a set of data points obtained as samples of a quantity, the common definition of a percentile, as introduced in elementary statistics, known as the nearest-rank method, is a value that splits a sorted data set into two parts: the lower part contains P percent of the data, and the upper part contains the rest of the data, i.e. $(100 - P)$ percent of the data. More formally, for a data set of N ordered values $x_1, \dots, x_N$ (sorted in ascending order), the P-th percentile (with $0 < P < 100$) is the smallest value x_h in the data set such that no more than P percent of the data is strictly less than x_h, i.e.

$$Prob\left(X < x_h\right) \le \frac{P}{100},$$

and at least P percent of the data is less than or equal to x_h, i.e.

$$Prob\left(X \le x_h\right) \ge \frac{P}{100}.$$

For a continuous random variable X,

$$Prob\left(X < x_h\right) = Prob\left(X \le x_h\right),$$

and so, the above definition reduces to

$$Prob\left(X \le x_h\right) = \frac{P}{100}.$$

Using the nearest-rank method, the P-th percentile is obtained by first calculating the ordinal rank h, and then taking the value from the ordered list that corresponds to that rank. The ordinal rank is calculated using the following formula:

$$h = \lceil Np \rceil,$$

where

$$p = \frac{P}{100}$$

and the symbol $\lceil \ \rceil$ represent the ceiling function, i. e. $\lceil y \rceil$ is the smallest integer that is $\ge y$.

Thus, using the nearest-rank method, the P-th percentile is computed as $Q_p = x_h$. Based on the nearest-rank method, a percentile is always a value in the original data set $x_1, \dots, x_N$.

A definition similar with nearest-rank method is used by the Vose software, but instead of assigning the lowest empirical cdf value x_h as the P-th percentile, it assigns the upper empirical cdf value in the ordered data set, i.e. $Q_p = x_{h+1}$.

The nearest-rank method provides good estimates when we need to compute a percentile from a known population, and when the number of values in the data set is at least 100. For data sets with fewer than 100 distinct values, the nearest-rank method can produce the same value for more than one percentile.

When we take samples from a population, and compute the percentile for each sample, the sample percentiles provide non-parametric estimators of the population percentile, based on a set of independent observations from the distribution. The distribution of the P-th sample quantile is known to be asymptotically normal around the P-th population quantile, with a variance that is not easily computed in practice (Stuart and Ord (1994)). While the P-th sample quantile will approach the P-th population quantile as the sample size increases, it is not known how quick the convergence process is (i.e., at what sample size on the asymptotic convergence the difference between the P-th sample quantile and P-th population quantile is below a desired precision level). For our purpose, the question of interest is how accurate the average of the sample percentiles over multiple repetitions is relative to the population percentile, when sampling is performed using different sample sizes.

There is a variety of ways to compute sample quantiles, and most of these methods use linear interpolation between adjacent ranks. There is also another class called the weighted percentile methods, which are not discussed here, because they are not as common as the linear interpolation methods. Since a theoretical quantile is the inverse of the theoretical cdf, an intuitive definition of a sample quantile is the inverse of the empirical cumulative distribution function (ecdf).

While in the nearest-rank method, the P-th percentile is determined by the x_h value in the ordered data set, as described above, the linear interpolation methods assign the P-th percentile as an interpolated value between two adjacent values in the ordered data set (usually between the x_h and x_{h+1} values), according to this formula

$$Q_p = (1-\gamma)x_j + \gamma x_{j+1},$$

Where j depends on the values of N and p (usually j is defined as $j = \lfloor Np + m \rfloor$, where the symbol $\lfloor \ \rfloor$ represent the floor function, i. e. $\lfloor y \rfloor$ is the greatest integer that is $\leq y$), y is the interpolation parameter (usually a fractional quantity obtained as $\gamma = Np - j$), and m is a method specific parameter ($0 \leq m \leq 1$), which determines how the method interpolates between two adjacent data points. Also, each method has a different way to estimate quantiles beyond the lowest and highest values in the sample (either by linear extrapolation, or assuming the end values of the sample).

Hyndman and Fan (1996) analyzed nine different methods of computing sample quantiles. Their goal was to advocate for a standard definition for percentiles so that percentiles obtained using different statistical software will be computed according to the same formula. While this goal hasn't been achieved yet, the Hyndman and Fan (1996) paper is the most widely cited article regarding this issue, and most statistical software packages use one of the nine methods described by Hyndman and Fan (1996). Next, these nine methods are presented, and also some of the software packages that use those methods as default are identified; most software packages have more than one method implemented, with one method being used by default, unless the user specifically select a different method (Analytica has only one of these methods implemented, Excel has 2 of these methods implemented, SAS has 5 of these methods implemented, and R has all 9 methods implemented).

Table 1 presents a summary of the nine sample quantile definitions from Hyndman and Fan (1996), reformatted from Wikipedia. In Table 1, the $\lfloor\ \rceil$ symbol indicates rounding to the nearest integer, choosing the even integer in case of a tie. The first three methods estimate the percentiles using a discrete rounding scheme, while the other six methods use a continuous interpolation scheme. All the methods used for computing the percentiles are non-parametric, which means that no distribution assumption is made for the population from where the sample is taken.

The differences between these methods are more apparent for small data sets, and also when there are large gaps between two adjacent sorted data values. For a large data set coming from a continuous probability distribution, the different definitions of percentiles will provide very similar results for most percentiles, but there can be larger differences between different methods for the extreme percentiles (either very small or very large percentiles). The first three methods use the ecdf to estimate the sample quantiles, while the last six methods use different variations of a piecewise-linear estimate of the cdf to determine the sample quantiles. When the ecdf estimation is used, a quantile is always an observed data value or the average of two adjacent data values. An advantage of the piecewise-linear methods is that since the inverse cdf is a continuous function, a small change to the probability value (such as from 0.9 to 0.91) results in a small change to the quantile estimates; by contrast, the first three methods do not have this property.

Table 1. Summary of the nine sample quantile definitions, from Hyndman and Fan (1996)[a].

Method	h	Q_p	Description
Type 1 (nearest-rank)	$Np + 1/2$	$x_{\lceil h - 1/2 \rceil}$	Inverse of the empirical cumulative distribution function.
Type 2 (default in SAS)	$Np + 1/2$	$(x_{\lceil h - 1/2 \rceil} + x_{\lfloor h + 1/2 \rfloor}) / 2$	Same as Type 1, but with averaging at discontinuities.
Type 3	Np	$x_{\lfloor h \rceil}$	The observation numbered closest to Np.
Type 4 (@Risk, Crystal Ball)	Np	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Linear interpolation of the empirical distribution function.
Type 5	$Np + 1/2$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Piecewise linear function where the knots are the values midway through the steps of the empirical distribution function.
Type 6 (Excel PERCENTILE.EXC)	$(N + 1)p$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Linear interpolation of the expectations for the order statistics for the uniform distribution on [0,1].
Type 7 (default in R, Analytica, Excel PERCENTILE)	$(N - 1)p + 1$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Linear interpolation of the modes for the order statistics for the uniform distribution on [0,1].
Type 8	$(N + 1/3)p + 1/3$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Linear interpolation of the approximate medians for order statistics.
Type 9	$(N + 1/4)p + 3/8$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	The resulting quantile estimates are approximately unbiased for the expected order statistics if data is normally distributed.

[a] https://en.wikipedia.org/wiki/Quantile

2.1. Percentile Computation Example

Figure 1 presents six of the nine methods from Table 1, plus the method used in the Vose software, for a small data set of 10 observations: {1, 2, 2, 3, 3, 4, 5, 5, 5, 9}. The 90th percentile computed using the seven methods ranges anywhere between the last two values in the data set, which are 5 and 9. The nearest-rank method assigns the lowest of the two values ($Q_{0.9} = 5$), while the Vose software assigns the highest of the two values ($Q_{0.9} = 9$). The default method in SAS (Type 2 method) assigns the middle of the two values ($Q_{0.9} = 7$), while the default method in R, which is also used by the Analytica software (Type 7 method), assigns the value $Q_{0.9} = 5.4$ as the 90-th percentile. Other values assigned by the other methods for the 90th percentile can vary anywhere between the last two values in the data set, as it can be seen in Figure 1 ($Q_{0.9} = 8.6$ for Type 6 method, $Q_{0.9} = 7.53$ for Type 8 method, and $Q_{0.9} = 7.4$ for Type 9 method). From the nine methods used in this example, three methods (Type 1, Type 2, and Vose) use a discrete rounding scheme, while the other methods use a continuous interpolation scheme. Not all the methods are displayed here, since some of them produce identical values for the percentiles; for the chosen data set and the 90th percentile, the Type 3 and Type 4 methods produce the same result as the Type 1 method, and the Type 5 method produce the same result as Type 2 method.

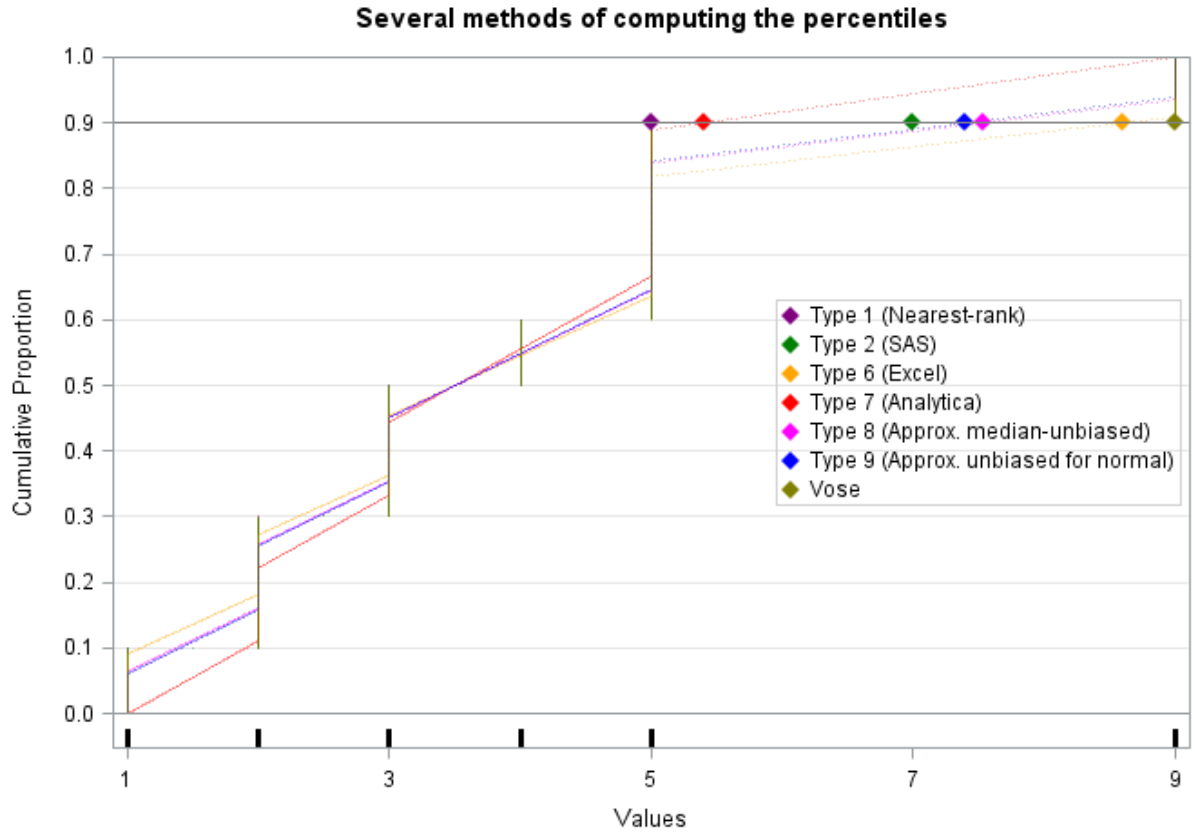


Figure 1. Several methods of computing the percentiles for a small data set.

2.2. Percentile Methods Literature Review

Parrish (1990) analyzed ten quantile estimators in small samples from a normal distribution with respect to the bias and mean squared error and concluded that the performance characteristics of individual estimators depend on the way in which the estimator is defined, the sample size, the quantile level being estimated, and the underlying parent distribution. The article also concludes that the larger the sample size, the less important will be the effect of the choice of estimator, and that estimates of the more extreme quantiles typically are subject to higher variability and bias as compared to those at intermediate levels such as the quartiles. Dielmann, Lowry, and Pfaffenberger (1994) extended the work by Parrish (1990), by analyzing small samples from skewed distributions (including the lognormal distribution), in terms of the mean squared error and mean absolute deviation and concluded too that there is not one percentile estimator that is always the best.

Hyndman and Fan (1996) evaluated their nine proposed methods against 6 desirable properties that a sample quantile should have. Only one of the nine methods (Type 5 method) satisfies all the 6 properties, but that method is not the method that Hyndman and Fan (1996) recommended, since it is a compromise definition in the sense that it is derived by interpolating between the midpoints of the inverse of the distribution function. Hyndman and Fan (1996) recommended Type 8 method to become the standard sample quantile definition across all the software packages, a goal that still had not been accomplished; the reason why Hyndman and Fan (1996) recommended Type 8 method, even though is satisfies only 5 of the 6 desirable properties, is because it gives approximately median-unbiased estimates of $Q(p)$, regardless of the distribution. Type 9 method, which also satisfies 5 of the 6 desirable properties, is very similar to Type 8 method, but is approximately unbiased only for the normal distribution, and not for other distributions. Type 7 method, which is used in the Analytica software, satisfies only 4 of the 6 desirable properties and is not an unbiased method.

Schoonjans, DeBacquer, and Schmid (2011) analyzed the accuracy of percentiles calculated in samples with sizes up to 1,000, from normal and lognormal distributed population data, and computed the 5^{th}, 25^{th}, 75^{th}, and 95^{th} percentiles for Type 2, Type 6, Type 7, and Type 8 methods. The article concluded that Type 2 method is the preferred one in general for continuous data and warned about the effect of the log-transformation on the calculation of percentiles, since the linear interpolation is applied in the calculation of percentiles, and the log-transformation changes the distribution model within the interpolated interval.

Kreutzmann (2018) investigated the bias and accuracy of different quantile definitions for income and wealth data; one simulation in this paper compared multiple quantile estimators, including all the nine methods described in Hyndman and Fan (1996), with the corresponding theoretical quantile from a Generalized Beta distribution (a right-skewed distribution with a very long tail). Kreutzmann (2018) analyzed the sample quantiles against their theoretical values, by taking repeated samples with sizes up to 1,000 values, and for the 99^{th} quantile, the article concluded that Type 5 is best method when looking at the smallest relative bias, while Type 7 method is best when looking at the smallest relative root mean squared error. The article also

concluded that the performance of the quantile estimators differs, depending on the quantile level, sample size, and type of distribution; that is, the sample size and quantile level at which the quantile estimators perform well in terms of bias and accuracy strongly depend on the type of distribution. Kreutzmann (2018) also points out that the larger the sample size, the closer the estimates from different quantile definitions are to each other, and the convergence of the quantile definitions is slower for the extreme quantiles than for the central quantiles.

Based on the above findings from literature, the search for the best quantile estimators in terms of bias and variability can be narrowed down to Type 2, Type 5, and Type 8 methods, which seem to be the best candidates for estimating the 99th percentile based on repeated samples from the same population. It's also clear that the larger the sample size, the more accurate each quantile estimator will be compared to the true theoretical value for each distribution. What still needs to be investigated is which of these methods is best to estimate the 99th percentile for the distributions most commonly used in NIOSH-IREP, and how do these methods compare with the current method implemented in NIOSH-IREP. A second question that needs to be addressed is what sample size should be used in the NIOSH-IREP Monte Carlo computations, so that the sample percentile estimates reach a desired level of precision (e.g., as compared to theoretical percentile values.)

3.0 PERCENTILE COMPUTATION SIMULATIONS

To get a better understanding of how some of the methods that are used to compute the percentiles in different software packages behave for the most common distributions used in NIOSH-IREP, two different simulations were set up. The first simulation involves sampling from a single distribution, and the second simulation involves combining values simulated from two distributions.

3.1. Percentile Computation Simulation for a Single Distribution

Five most common continuous distributions used in the NIOSH-IREP software were tested, each with different parameters settings, for a total of 10 combinations: Uniform (1,6), Triangular (1,3,5), Triangular (1,5,5), Weibull (2,1), Weibull (1,10), Normal (5,1), Normal (25,5), Lognormal (1,3), Lognormal (3,6), and Lognormal (3,10).

Different sampling methods can be used to take random samples from a known distribution. The most common methods are the Median Latin Hypercube (MLH), the Random Latin Hypercube (RLH), and the Monte Carlo (MC) sampling. The MLH method is the default method used by the Analytica engine, and the NIOSH-IREP software is using this method of random sampling to

generate all probability distributions used in the computational process of generating PC values. In this paper, only the MLH and RLH sampling method were used, each with a different number of iterations (i.e., multiple iterations, or repetitions were used to assess the distribution of the percentiles tested). The Monte Carlo sampling method was not used in this simulation because the range of values obtained is too large, compared with the MLH and RLH sampling methods.

To analyze the distribution of values for each percentile definition, multiple repetitions were used to take samples from each distribution. For the RLH sampling, $R = 500$ repetitions were used to assess the distribution of the percentiles computed from the samples. Since for the MLH sampling, the resulting samples of the same size always have the same values when they are sorted, and as a result the percentiles are the same for all MLH samples of the same size, only one repetition was performed for the MLH sampling.

Six different sample sizes were used: 2,000, 5,000, 10,000, 20,000, 30,000, and 50,000. The default sample size currently used in NIOSH-IREP is 2,000, and a 10,000 sample size is used in NIOSH-IREP Enterprise Edition, where 30 runs are performed, and results are averaged to determine the final PC value.

Six different methods of computing the percentiles were chosen for this simulation, including the methods implemented in the most common software packages: Type 1 method (Nearest-Rank method), Type 2 (default method used in the SAS software), Type 4 (used in the @Risk software), Type 6 (used by the Excel PERCENTILE.EXC function), Type 7 (used in the Analytica software), and the method used by the Vose software.

Figure 2 shows the box plots for the 50^{th} percentile computed based on 6 different methods, using both MLH and RLH sampling, at different sample sizes, sampled from the Lognormal LN(3,6) distribution. Each of these boxplots shows the first quartile, the median, and the third quartile of these distributions, while the whiskers from each boxplot extend to the entire range of the sampled values. The median in each boxplot is indicated by a line, while the mean is displayed as a diamond symbol in the boxplots. Three of the software packages (Excel, Analytica, and SAS) show very little bias in estimating the 50^{th} percentile at each sample size. Two other methods (Nearest-Rank and @Risk) underestimate the 50^{th} percentile computed from the samples when compared with the theoretical 50^{th} percentile of the distribution, and the Vose software overestimates the 50^{th} percentile. The results obtained using the MLH sampling method match very well with the median result obtained from the RLH sampling method, but the box plot for the distribution of the RLH sampling show that the 50-percentile estimate can vary a lot from sample to sample.

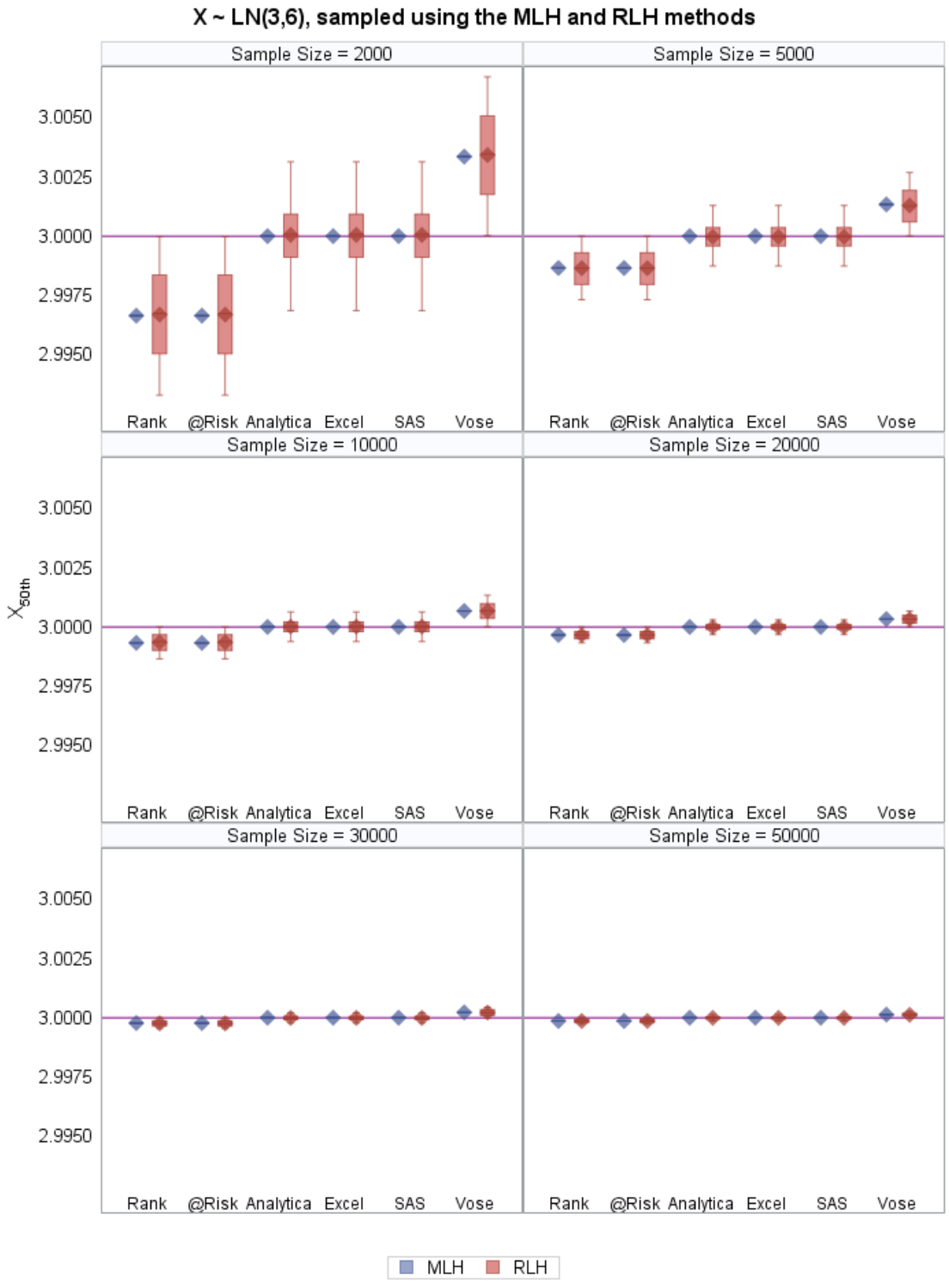


Figure 2. The 50th percentile computed using 6 different methods, for the LN(3,6) distribution.

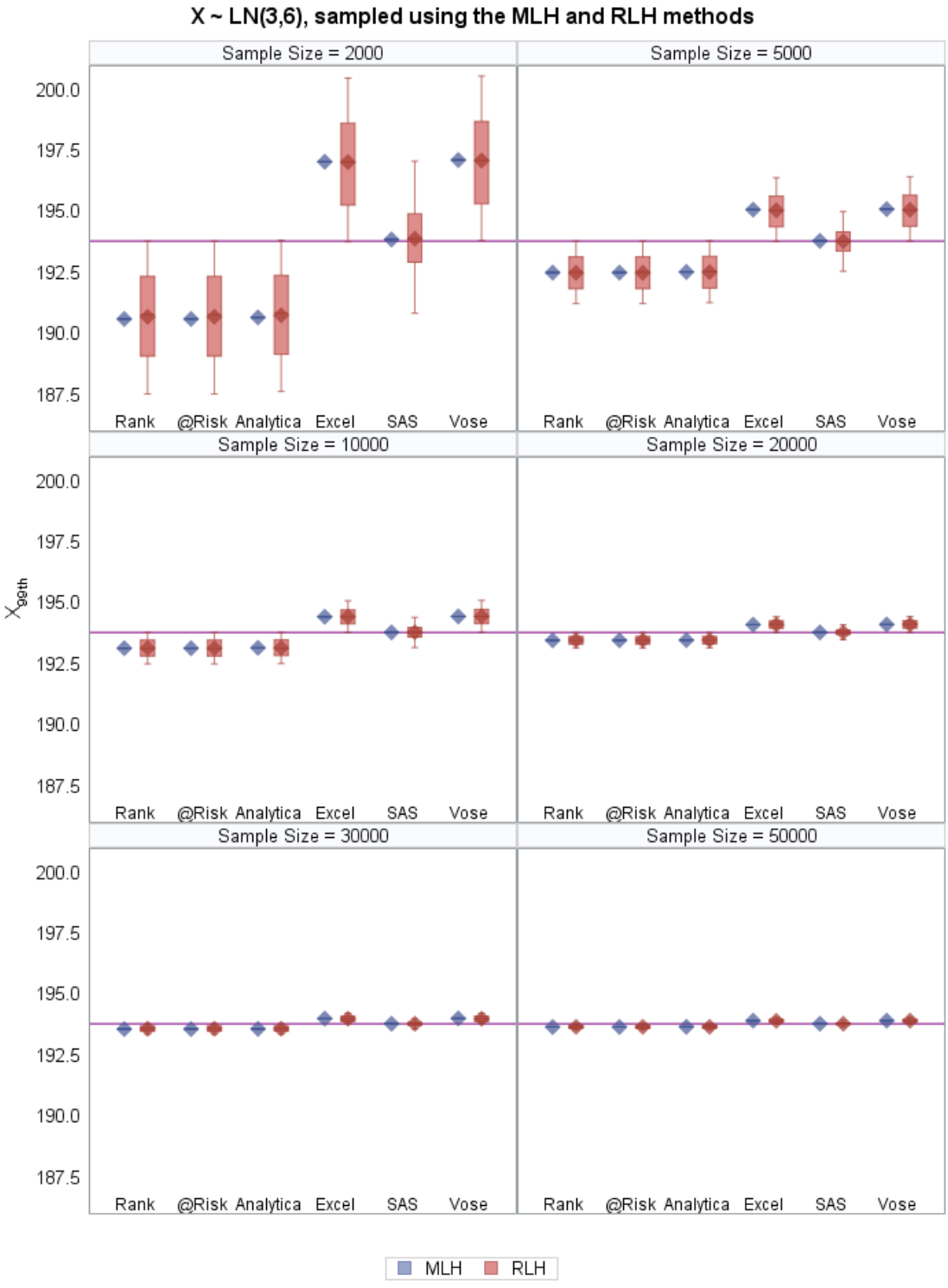


Figure 3. The 99th percentile computed using 6 different methods, for the LN(3,6) distribution.

Figure 3 shows the box plots for the 99th percentile computed based on 6 different methods, using MLH and RLH sampling, at different sample sizes, sampled from the LN(3,6) distribution. The method used by the SAS software has the smallest bias in estimating the 99th percentile at each sample size. Three of the methods used (Nearest-Rank, @Risk, and Analytica) underestimate the 99th percentile computed from the samples when compared with the theoretical 99th percentile of the distribution, and two methods (Excel, and Vose) overestimate the 99th percentile. Similarly, as for the 50th percentile, the results obtained using the MLH sampling method match very well with the median result obtained from the RLH sampling method, but the box plot for the distribution of the RLH sampling show that the 99-percentile estimate can vary a lot from sample to sample.

For both the 50th percentile, as well as the 99th percentile, the estimate based on the MLH sampling, as well as the average estimate from the RLH sampling, get closer to the theoretical percentile, as the sample size increases.

The results obtained for the 99th percentile from all distributions tested in the simulation are shown in Appendix A. Among all methods tested, Type 2 method used as default in the SAS software seems to be the one with the smallest bias at all the sample sizes, and for all the simulation scenarios tested.

As expected, the estimates of the 99th percentile obtained based on a sample of 2,000 are far off compared with the estimates obtained based on a larger sample, such as 20,000 iterations or more. To compare how far the estimates obtained for different sample sizes are from the theoretical population quantile, the bias is computed for each method.

The absolute bias (B) for each definition Q and each quantile level p is defined as

$$B = \frac{1}{R}\sum_{r=1}^{R}(\widehat{Q}_{p,r} - Q_p),$$

Where R is the number of repetitions, $\widehat{Q}_{p,r}$ is the quantile estimate in repetition r, and Q_p is the value of the theoretical population quantile.

Table 2 shows the absolute bias at three different sample sizes (2,000, 10,000, and 20,000 repetitions), based on the method used in the Analytica software of estimating the 99th percentiles, for the 10 simulated scenarios. The bias reported in Table 2 is obtained using the MLH method, but the RLH method produces almost identical results. For all the distributions tested at all sample sizes, Type 7 method used by Analytica to estimate percentiles produces negative biases, meaning that this method is underestimating the theoretical quantiles. Skewed distributions with the largest uncertainty have the largest bias, and the bias decreases at the sample size increases.

Table 2. Bias at different sample sizes, based on MLH sampling and the method used in the Analytica software for the estimation of the 99th percentiles.

Distribution	Bias at 2,000 sample size	Bias at 10,000 sample size	Bias at 20,000 sample size
Uniform (1,6)	-0.0012	-0.0002	-0.0001
Triangular (1,3,5)	-0.0034	-0.0007	-0.0003
Triangular (1,1,5)	-0.0049	-0.0010	-0.0005
Weibull (2,1)	-0.0056	-0.0011	-0.0006
Weibull (1,10)	-0.2419	-0.0489	-0.0245
Normal (5,1)	-0.0091	-0.0018	-0.0009
Normal (25,5)	-0.0455	-0.0092	-0.0046
LogNormal (1,3)	-0.1280	-0.0259	-0.0130
LogNormal (3,6)	-3.1305	-0.6359	-0.3186
LogNormal (3,10)	-13.1705	-2.6806	-1.3433

Another way to compare methods for computing percentiles is to look at the relative bias, and the relative root mean squared error for each method. The goal is to identify the method with the smallest relative bias and the smallest relative root mean squared error, as such method provides estimates closest to the true quantile level and produces the least variability over repeated sampling.

The relative bias (*RB*) for each definition Q and each quantile level p is defined as

$$RB = \frac{1}{R}\sum_{r=1}^{R}\frac{(\widehat{Q}_{p,r} - Q_p)}{Q_p},$$

Where R is the number of repetitions, $\widehat{Q}_{p,r}$ is the quantile estimate in repetition r, and Q_p is the value of the theoretical population quantile.

The relative root mean squared error (*RRMSE*) for each definition Q and each quantile level p is defined as

$$RRMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\widehat{Q}_{p,r} - Q_p}{Q_p}\right)^2}.$$

Figure 4 shows the relative bias of the 99th percentile computed based on 10 different methods (the 9 methods from Hyndman and Fan (1996), and the method implemented in Vose software), at different sample sizes, sampled from the LN(3,6) distribution, and using the RLH sampling. The relative bias of the 99th percentile computed using the MLH sampling (not shown) is almost identical with the relative bias from the RLH sampling.

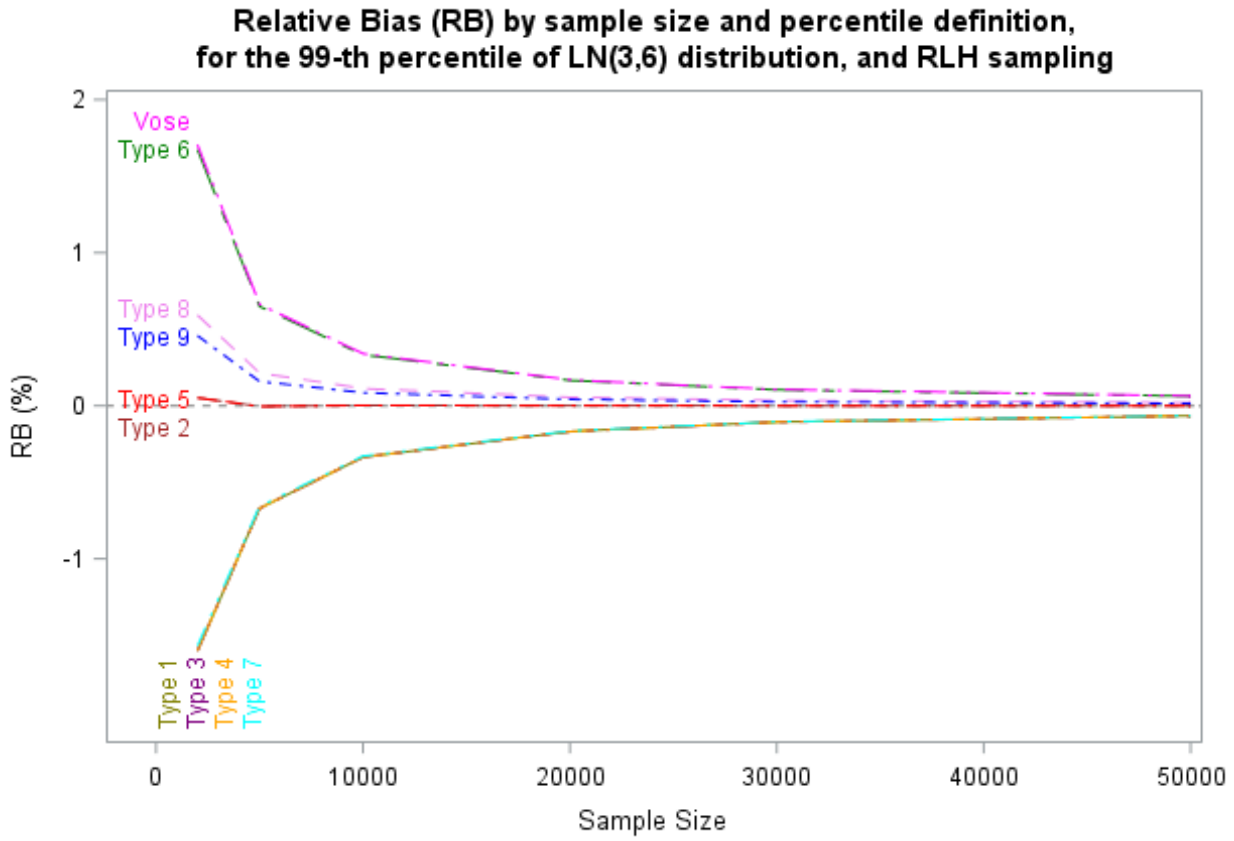


Figure 4. Relative Bias for the 99th percentile computed for 10 different percentile estimation methods.

Figure 5 shows the relative root mean squared error of the 99th percentile computed based on 10 different methods, at different sample sizes, sampled from the LN(3,6) distribution, and using the RLH sampling. Figure 6 shows the relative root mean squared error of the 99th percentile computed based on 10 different methods, at different sample sizes, sampled from the LN(3,6) distribution, and using the MLH sampling.

The relative bias and the relative root mean squared error of the 99th percentile sampled from the other distributions are not shown, but they are very similar to the ones for the LN(3,6) distribution.

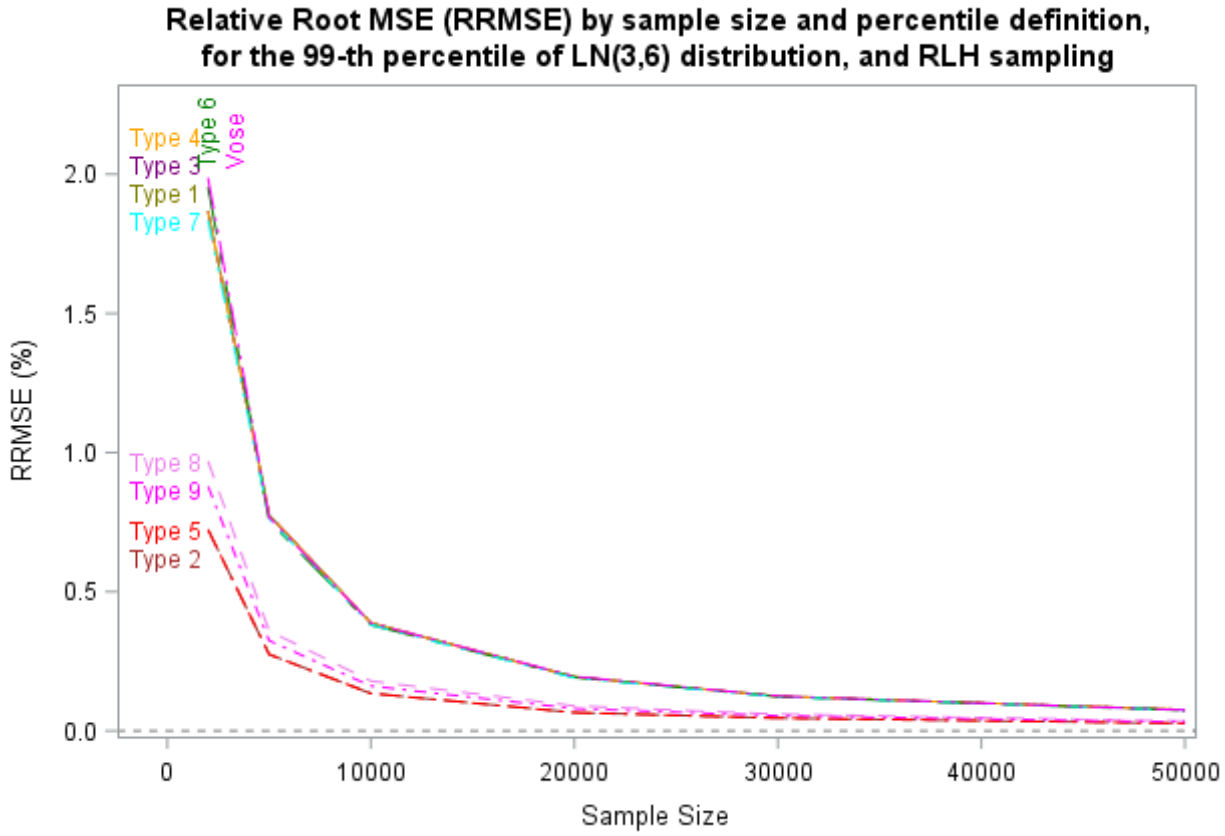


Figure 5. RRMSE for the 99th percentile computed for 10 percentile estimation methods, and RLH sampling.

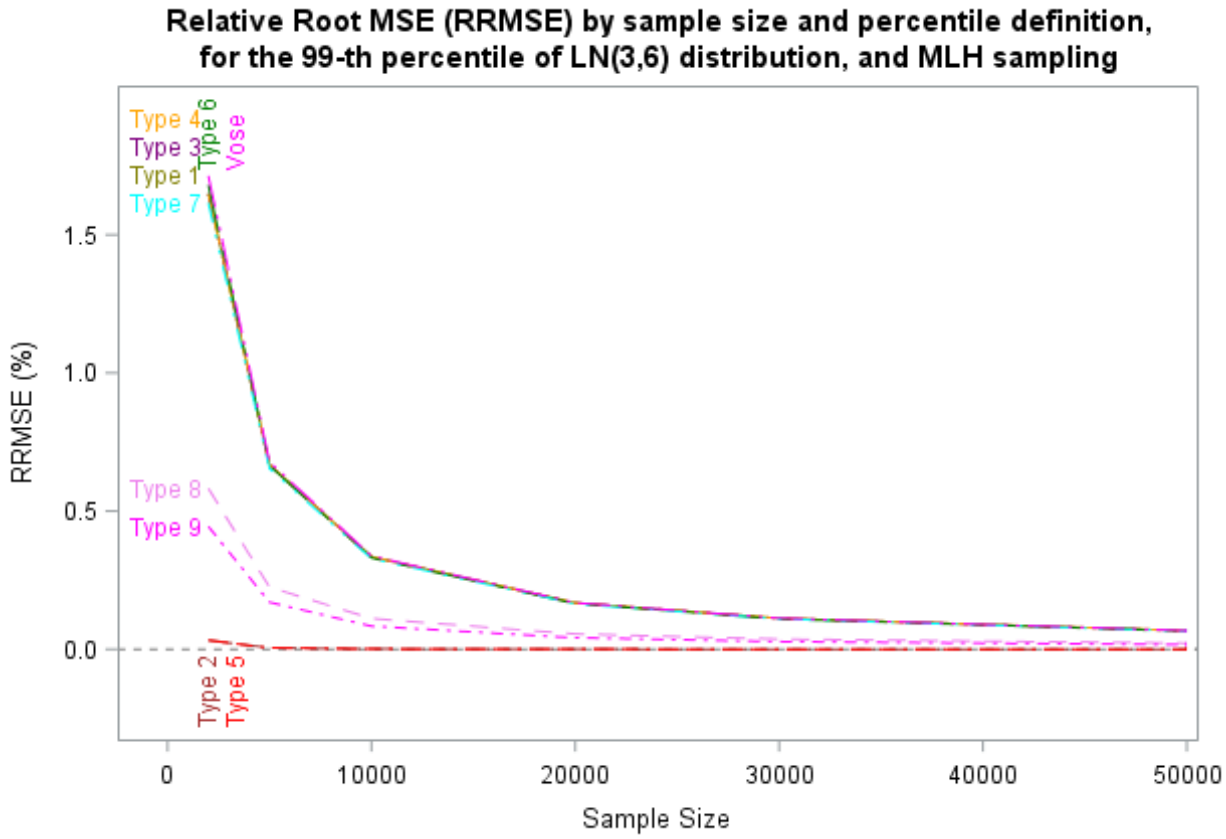


Figure 6. RRMSE for the 99th percentile computed for 10 percentile estimation methods, and MLH sampling.

In Figures 4, 5, and 6, some of the methods have identical or very close values for the relative bias and the relative root mean squared error, the reason being that some of methods used to compute the 99th percentile produced the same estimates (for example Type 2 and Type 5 methods produce identical results in estimating the 99th percentiles for the sample sizes selected in this simulation). With respect to both relative bias and relative root mean squared error, Type 2 and Type 5 methods perform very well, having the lowest values at all sample sizes. As the sample size increases, both the relative bias and the relative root mean squared error decrease, and the differences between the methods become much smaller. The MLH sampling seems to have an advantage over the RLH sampling in terms of the RRMSE, but for the relative bias both sampling methods produce similar results.

In conclusion, among the tested methods used to compute the 99th percentile in this simulation, Type 2 and Type 5 methods are the ones with the smallest relative bias and smallest relative root mean squared error at all the sample sizes, and for all the distributions scenarios tested, and the MLH sampling produces a lower RRMSE compared to RLH sampling.

3.2. Percentile Computation Simulation for Combinations of Two Distributions

To investigate which method of computing the 99th percentile is better in terms of bias and variability when distributions are combined in NIOSH-IREP, a second simulation was set up.

Since the normal and lognormal distributions are the most common distributions used in NIOSH-IREP, these distributions, with either small or large standard deviations, were used in this simulation, for both additive and multiplicative interactions. A mixed interaction between the normal and lognormal distributions was also analyzed. A total of 20 scenarios were considered (Table 3), obtained by either adding or multiplying two distributions from the following list: Normal (5,1), Normal (5,25), Lognormal (1,3), and Lognormal (3,6).

When two normal distributions are added, or when two lognormal distributions are multiplied, the distribution obtained is known to be either normal or respectively lognormal, and its parameters can be determined based on the parameters of the two input distributions. Therefore, in 6 out of the 20 scenarios, the theoretical 99th percentile of the combined distribution could be easily determined. For the remaining 14 scenarios, where the combined distribution is not known, the theoretical 99th percentile of the combined distribution cannot be determined exactly, however a confidence interval can be computed for the estimated 99th percentile.

There are several methods of computing confidence intervals for the percentiles, either assuming that the data is normally distributed, or making no assumption regarding how the data is distributed. For the 14 scenarios when the combined distribution is not known, distribution-free confidence intervals were computed and reported in Table 3 for the 99th percentile of the combined distributions. The confidence intervals for the 99th percentile were computed according to the distribution-free method described by Hahn and Meeker (1991). In computing the distribution-free confidence intervals (CIs), the only assumption is that the sampling is done randomly from the population.

Table 3. Theoretical 99th percentile estimates, or the 95% CIs for the 20 scenarios tested in combining two distributions.

Distr. #1	Distr. #2	(Distr. #1) + (Distr. #2) 99th percentile estimate, or 95% CI for the 99th percentile	(Distr. #1) * (Distr. #2) 99th percentile estimate, or 95% CI for the 99th percentile
N(5,1)	N(5,1)	13.29	(43.70, 43.72)
N(25,5)	N(25,5)	66.45	(1092.72, 1093.22)
N(5,1)	N(25,5)	41.86	(218.52, 218.62)
LN(1,3)	LN(1,3)	(18.96, 19.00)	37.13
LN(3,6)	LN(3,6)	(315.56, 316.60)	3268.31
LN(1,3)	LN(3,6)	(195.36, 196.07)	398.62
N(5,1)	LN(1,3)	(17.99, 18.02)	(65.77, 65.92)
N(5,1)	LN(3,6)	(198.45, 199.17)	(973.15, 976.82)
N(25,5)	LN(1,3)	(41.56, 41.59)	(328.81, 329.56)
N(25,5)	LN(3,6)	(218.61, 219.32)	(4864.43, 4882.89)

Since the distribution of the sum or product of two distributions is not always known, it needs to be simulated through a Monte Carlo process, and this can be done using any of the three sampling methods, either the MLH, the RLH, or the MC sampling methods. For each of the 14 scenarios when two distributions were either added or multiplied, and the combined distribution was not known, a separate simulation with 50,000,000 iterations was performed for each of the three sampling methods; the reason why such a large sample size was used is that as the sample size gets larger, the width of the confidence interval gets shorter, thus providing a very accurate range of values for the theoretical 99th percentile. The distribution-free 95% CIs obtained by sampling from the combined distribution, through each of the three sampling methods, are very similar in range, but the confidence intervals obtain through the RLH method seems to provide the closest point estimate to the theoretical 99th percentile most of the time, for the 6 scenarios when the 99th theoretical percentile is known; for this reason, the distribution-free 95% CIs obtained based on the RLH sampling are reported in Table 3 (the computations of the 95% CIs was performed using the SAS software).

Once the theoretical 99th percentiles or their 95% CIs were determined for each of the 20 scenarios, the next step was to perform a new simulation to determine which method of computing the 99th percentile based on repeated sampling is better in terms of bias and variability for the combined distributions in the 20 scenarios. For each of these 20 scenarios, the two distributions were combined using three different sampling methods: MLH, RLH, and MC methods.

Different sample sizes were used in this simulation, to assess when the convergence to the theoretical percentile is achieved. The set of sample sizes used is N={2000, 3000, 5000, 7000, 10000, 20000, 30000, 50000, 70000, 100000}. For a given sample size, different random seeds are used to generate the random values from the input distributions, to produce the combined distribution. This process was repeated either 500 times or 1000 times, depending on the sample sizes used for the input distributions. For sample sizes N<10,000, the number of repetitions used was $R = 1000$, and for sample sizes N≥10,000, the number of repetitions used was $R = 500$, since the results are more stable, and there is no need to repeat the sampling procedure so many times, and also because the whole simulation process uses a very large amount of computer resources.

In the previous simulation, when different methods of computing the percentiles were tested for one distribution only, it was concluded that the Type 2 and Type 5 methods of computing the 99th percentile seems to be the ones with the smallest relative bias and smallest relative root mean squared error at all the sample sizes. For this simulation where two distributions are combined, four different methods of computing the percentiles were analyzed in terms of bias and variability: Type 2 (default method used in the SAS software), Type 5 (the only unbiased method, satisfying all the 6 desired criteria in Hyndman and Fan (1996)), Type 7 (used in the Analytica software), and Type 8 (the method recommended by Hyndman and Fan (1996)). For each of the 20 scenarios, the 99th percentiles computed using the Type 2 and Type 5 methods produced the same results for all the sample sizes, and for each of the three sampling methods.

To compute the relative bias and the relative root mean squared error for the different estimators of the 99th percentile, the 99th theoretical percentile need to be known, but for the 14 scenarios where the theoretical 99th percentile is not known, it was assumed to be the midpoint of the 95% confidence interval, as reported in Table 3.

Figures 7 and 8 show the relative bias of the 99th percentile computed based on 3 different methods (Type 2, Type 7, and Type 8), at different sample sizes, sampled from the Y1= X1 + X2, and respectively Y2 =X1 * X2, where X1 ~ LN(1,3), and X2 ~ LN(3,6), and both Y1 and Y2 were obtained using the MLH sampling. The relative bias of the 99th percentile computed using the RLH sampling (not shown) is almost identical with the relative bias from the MLH sampling. The Type 2 method seems to perform better overall in terms of relative bias, at most sample sizes for both Y1 and Y2 distributions.

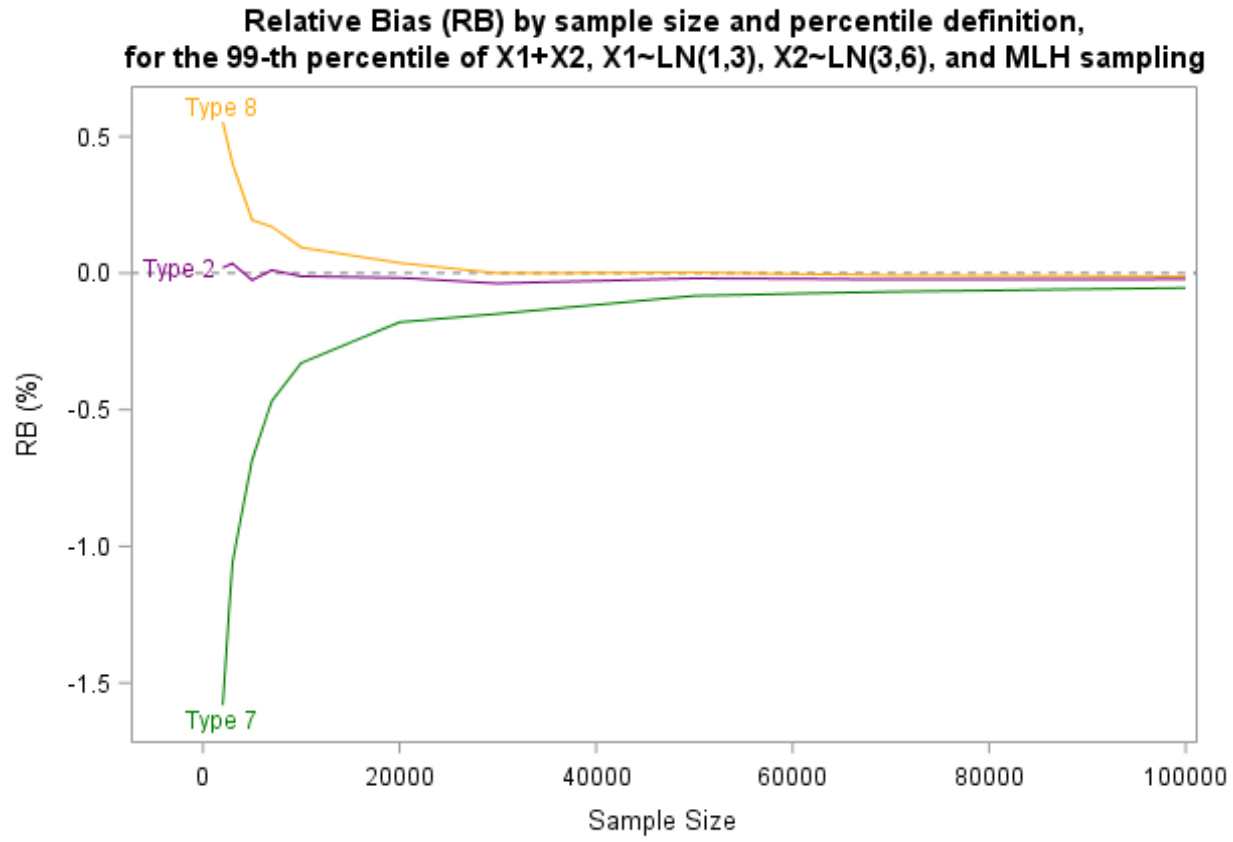


Figure 7. Relative Bias for the 99th percentile for X1 + X2, where X1 ~ LN(1,3), and X2 ~ LN(3,6).

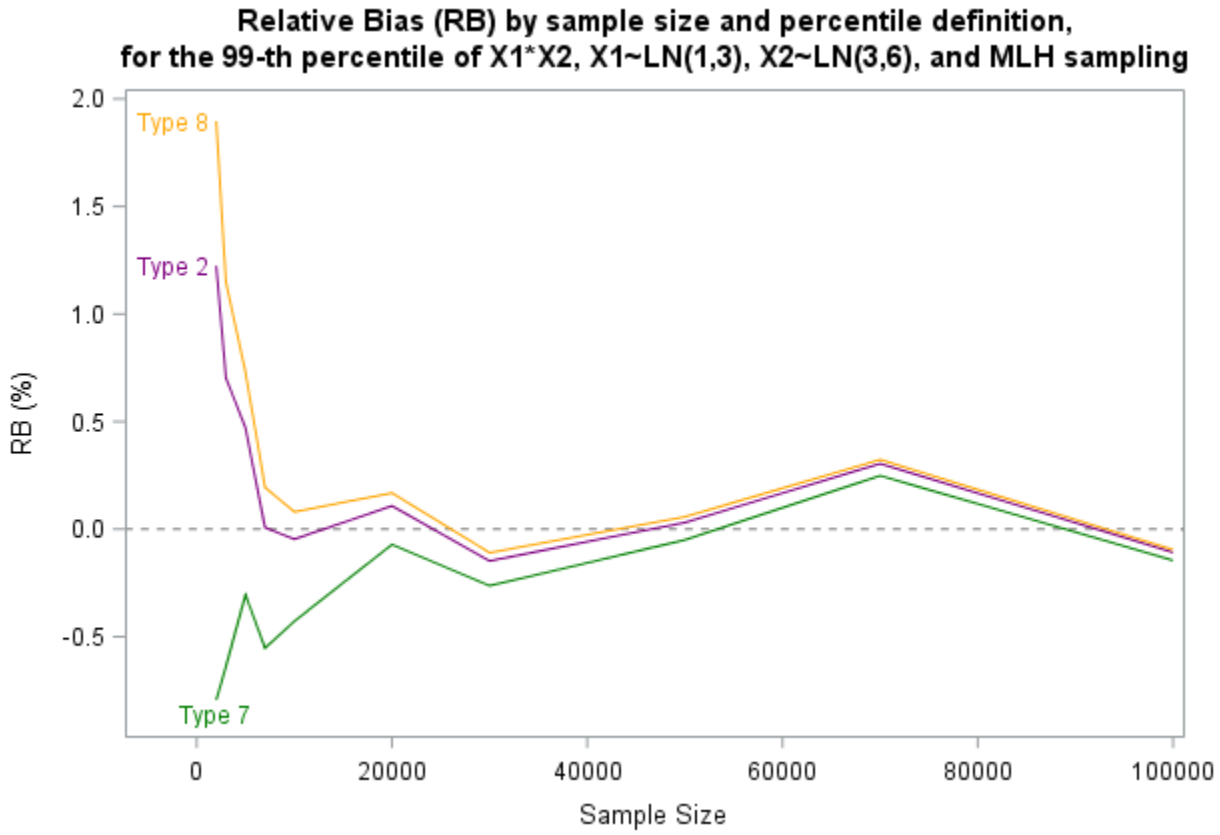


Figure 8. Relative Bias for the 99th percentile for X1 * X2, where X1 ~ LN(1,3), and X2 ~ LN(3,6).

Figures 9 and 10 show the relative root mean squared error of the 99th percentile computed based on 3 different methods (Type 2, Type 7, and Type 8), at different sample sizes, sampled from Y1= X1 + X2, and respectively Y2 =X1 * X2, where X1 ~ LN(1,3), and X2 ~ LN(3,6), and both Y1 and Y2 were obtained using the MLH sampling. The relative root mean squared error of the 99th percentile computed using the RLH sampling (not shown) follows the same pattern but is slightly higher than the relative root mean squared error from the MLH sampling. For the Y1 distribution, the Type 2 method performs better at smaller sample sizes in terms of relative root mean squared error, with all three methods converging as the sample size get larger. For the Y2 distribution, all three methods perform the same in terms of relative root mean squared error.

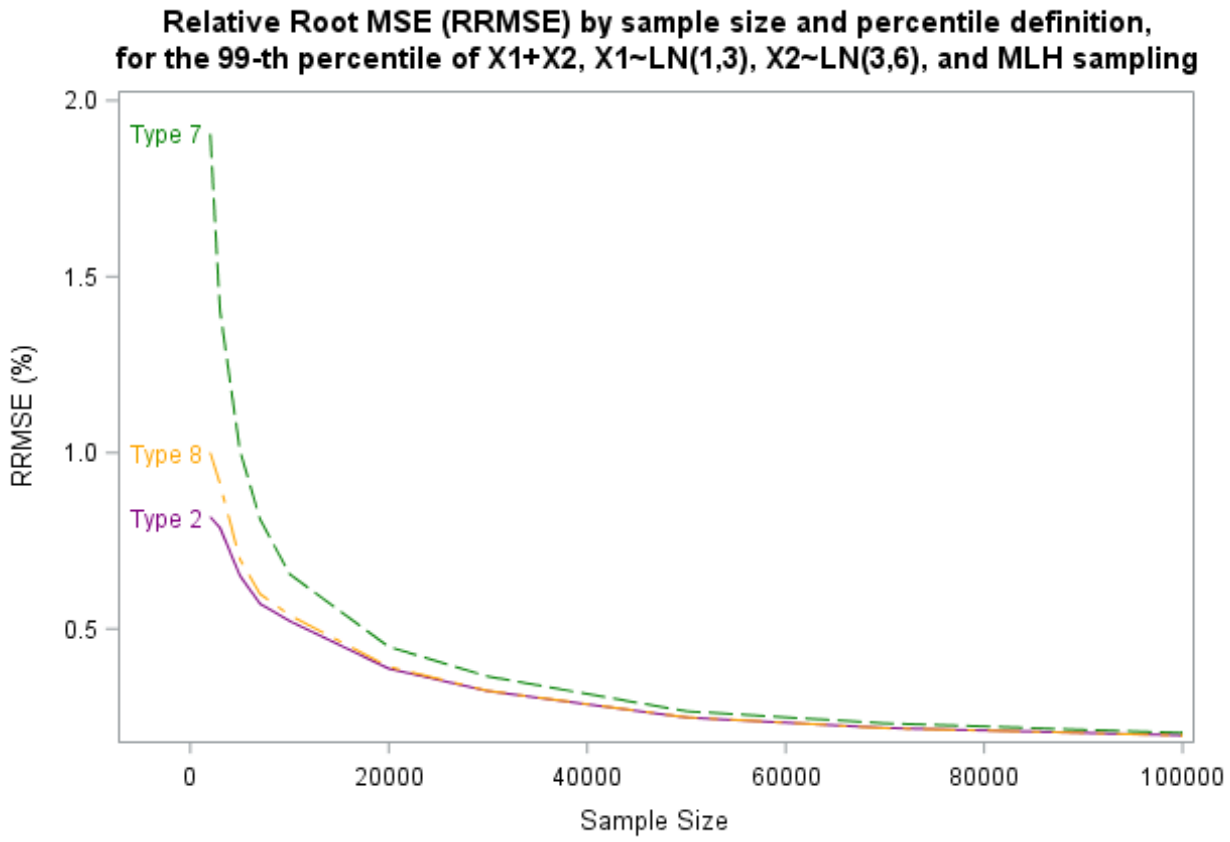


Figure 9. RRMSE for the 99th percentile computed for X1 + X2, where X1 ~ LN(1,3), and X2 ~ LN(3,6).

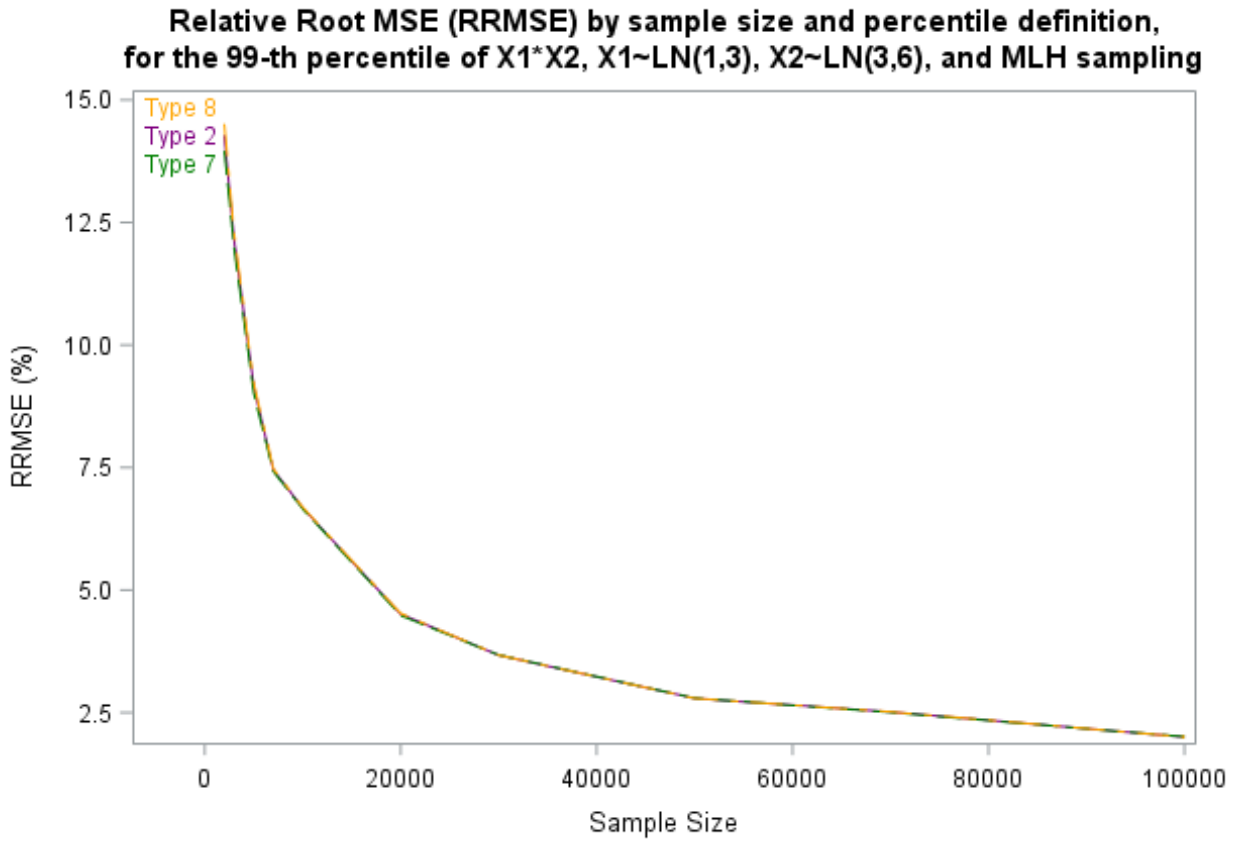


Figure 10. RRMSE for the 99th percentile computed for X1 * X2, where X1 ~ LN(1,3), and X2 ~ LN(3,6).

Figures 11 and 12 show the 99th percentile computed based on 2 different methods (Type 7 method used in Analytica, and Type 2 method used by default in SAS), at different sample sizes, sampled from the Y1= X1 + X2 distribution, where X1 ~ LN(1,3), and X2 ~ LN(3,6), and for all the three sampling methods. The blue dots represent the average of the 99th percentile over the number of repetitions indicated before, and the vertical red lines extend to one standard error around the average estimate. The red dotted horizontal lines in each plot represent the 95% confidence limits for the 99th percentile of the Y1= X1 + X2 distribution.

Figures 13 and 14 show the 99th percentile computed based on 2 different methods (Type 7 method used in Analytica, and Type 2 method used by default in SAS), at different sample sizes, sampled from the Y2= X1 * X2 distribution, where X1 ~ LN(1,3), and X2 ~ LN(3,6), and for all the three sampling methods. The red horizontal line in each plot represents the theoretical 99th percentile of the Y2= X1 * X2 distribution.

The method used by the SAS software has the smallest bias in estimating the 99th percentile at each sample size. The method used in Analytica underestimates the 99th percentile computed from the samples when compared with the theoretical 99th percentile of the distribution, at all sample sizes up to 20,000 iterations.

The average estimates obtained through the MLH and RLH sampling are similar for the Y1 distribution, but the average estimate obtained through the MC sampling can be quite different and can be outside the 95% confidence interval; also, the standard error obtained through the MC sampling is very large, compared with the standard errors obtained through the MLH and RLH methods. The average estimates obtained through all the sampling methods are much closer for the Y2 distribution, with the standard error obtained through the MC sampling being slightly larger than the standard errors obtained through the other two sampling methods.

The results obtained for the 99th percentile computed based on 2 different methods, for the all the 20 scenarios listed in Table 3, are shown in Appendix B. From the methods tested used to compute the percentiles in this simulation, the default method used by the SAS software seems to be the one with the smallest bias at all the sample sizes, and for all the simulation scenarios tested.

In conclusion, from the tested methods used to compute the 99th percentile in this simulation, the Type 2 method used by default in SAS seems to be the ones with the smallest relative bias and the MLH sampling seems to have an advantage over the RLH sampling in terms of the RRMSE for some of the scenarios tested. The Type 7 method used in Analytica underestimates the theoretical 99th percentile of the combined distribution in all scenarios, for most of the sample sizes, and especially for sample sizes up to 20,000 values the bias can be significant.

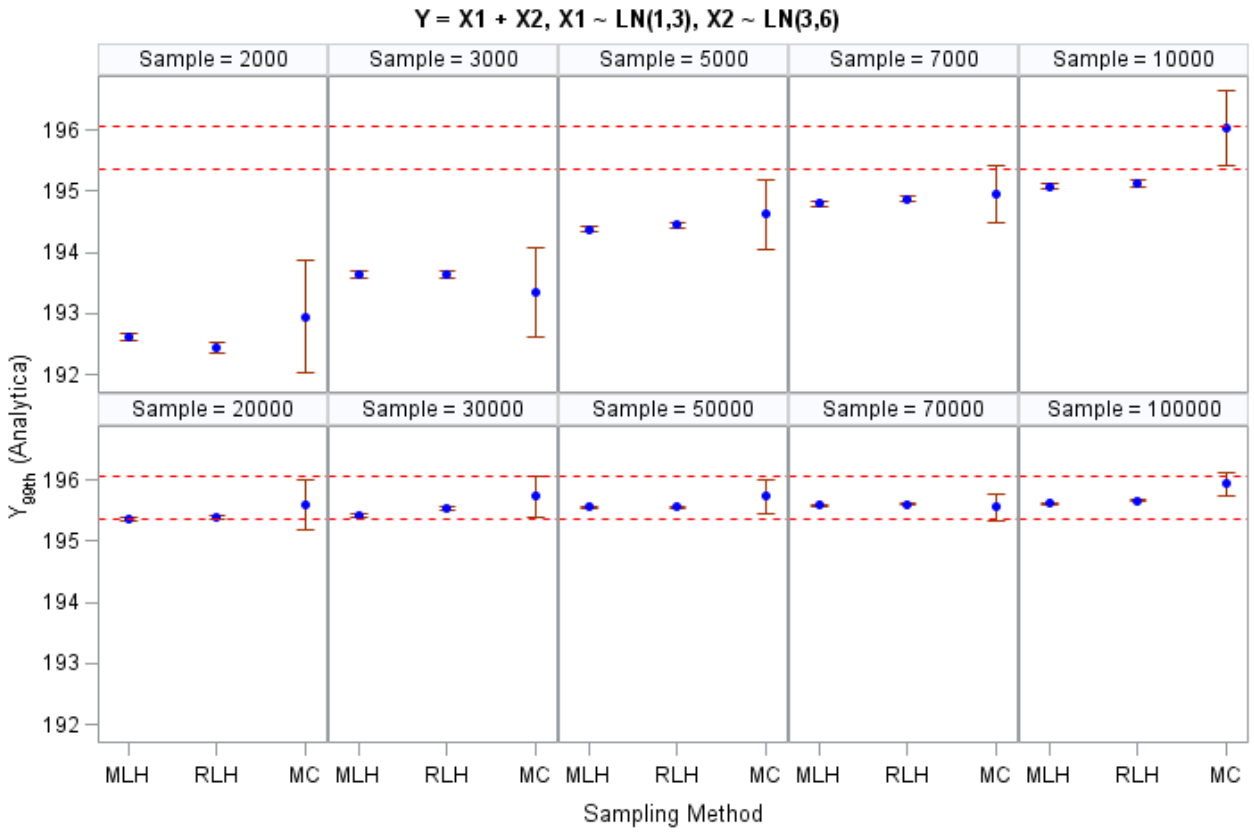


Figure 11. The 99th percentile computed using Type 7 method, for the sum of two distributions.

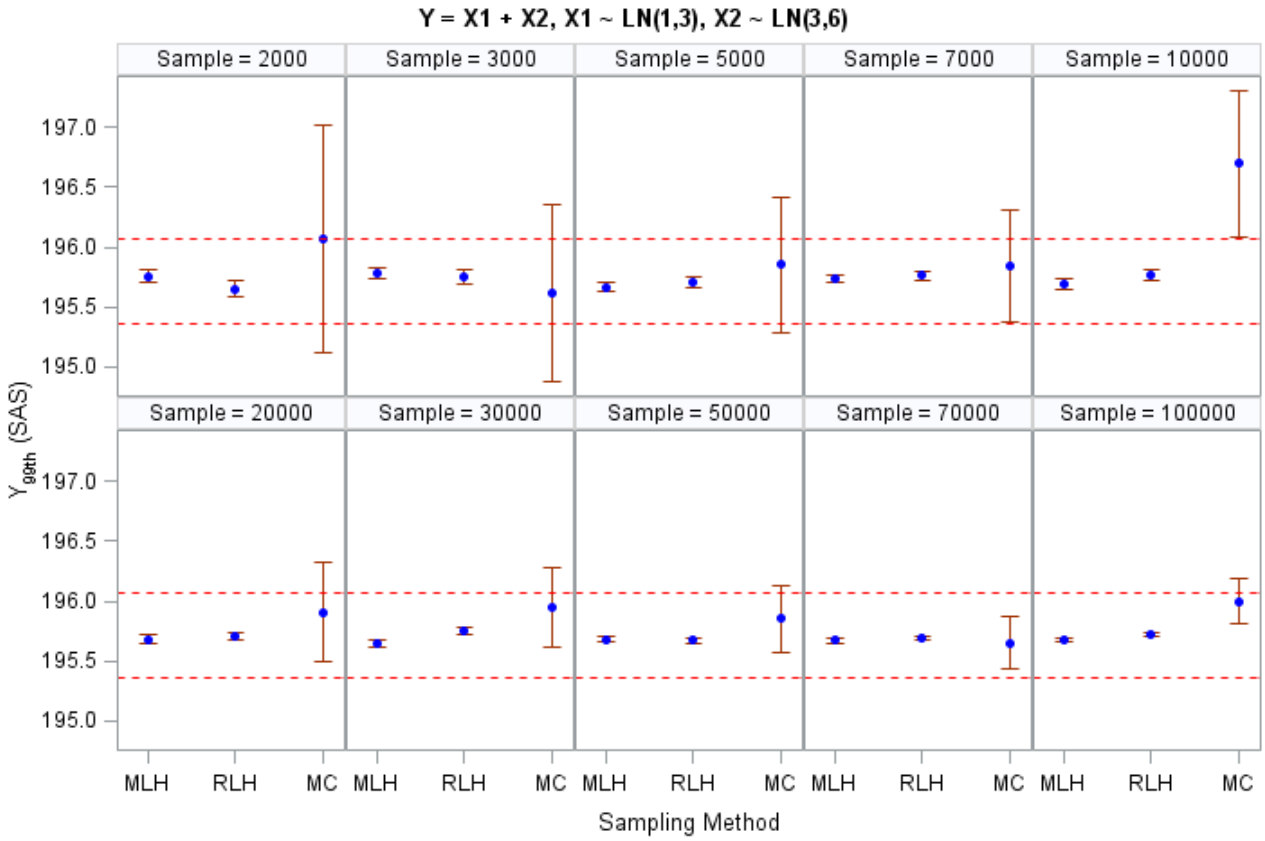


Figure 12. The 99th percentile computed using Type 2 method, for the sum of two distributions.

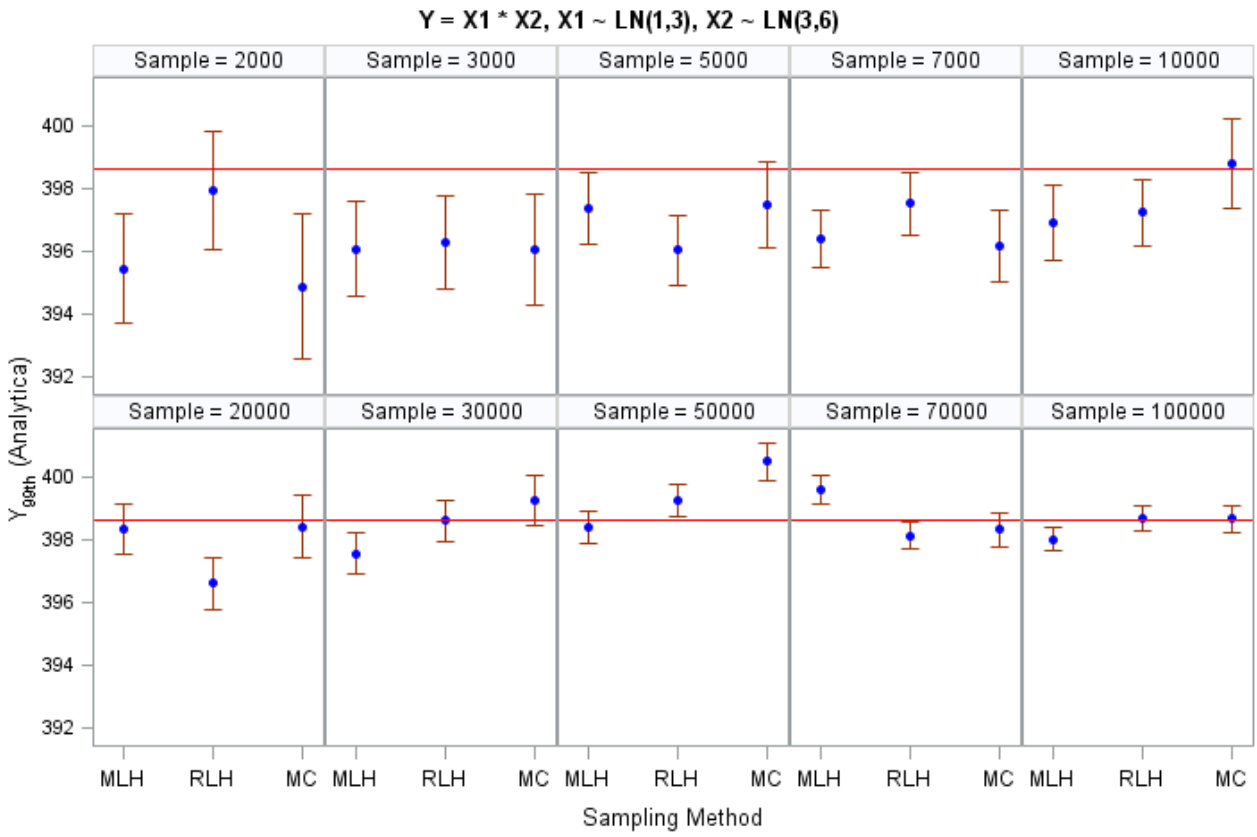


Figure 13. The 99th percentile computed using Type 7 method, for the product of two distributions.

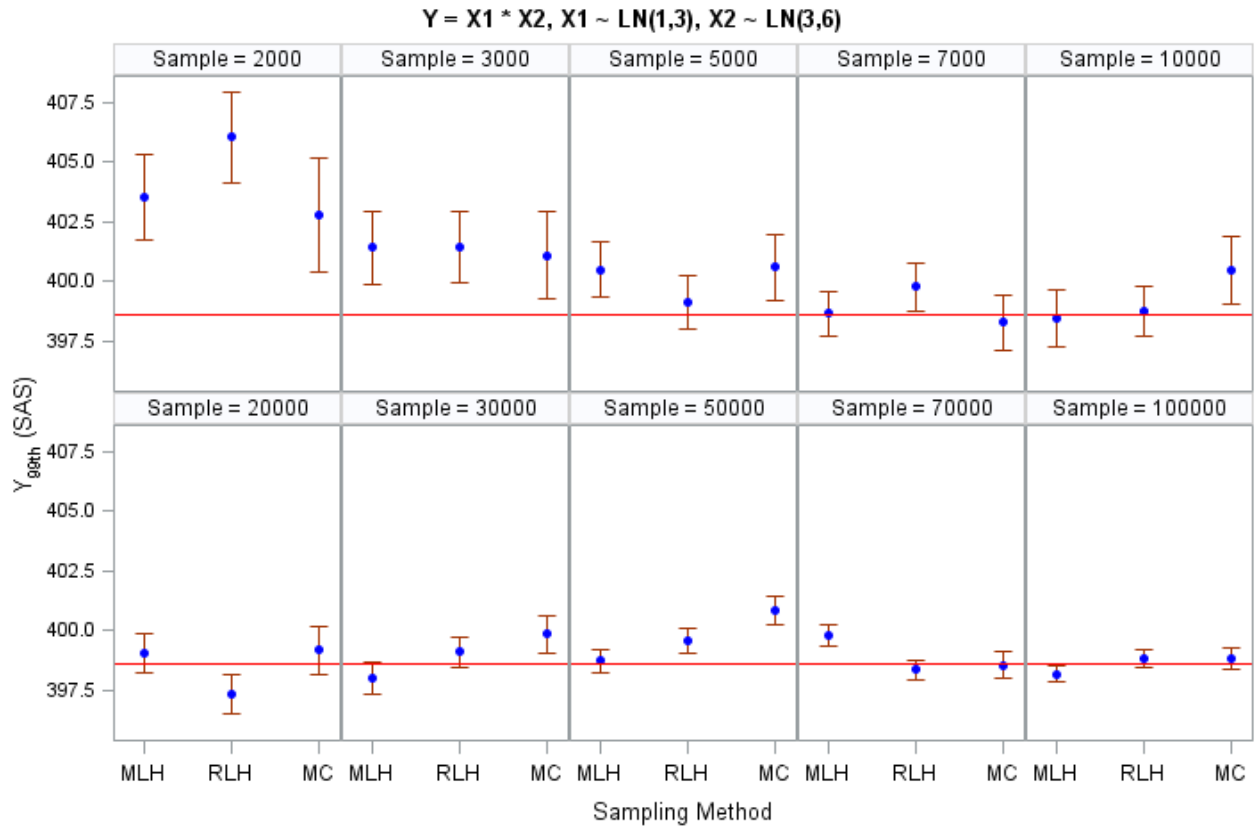


Figure 14. The 99th percentile computed using Type 2 method, for the product of two distributions.

4.0 CONCLUSIONS

There is no standard method to estimate a percentile (quantile) given multiple samples from a probability distribution. There are multiple definitions currently in use, with different software packages using different methods to compute percentiles. The performance of a percentile method differs, depending on the percentile level, sample size, type of distribution, magnitude of uncertainty and of positive skewness; however, the larger the sample size, the less important the effect of the percentile definition used is. This document summarizes and compares several percentile definitions and identifies which of these definitions is used in the NIOSH-IREP software for the purpose of computing PC values.

The document provides estimates of the 99th percentile, based on sampling from one distribution, and based on sampling from the sum or product of two most common distributions used in the NIOSH-IREP software.

For all the simulations scenarios tested, the Type 2 method, used by default in the SAS software, seems to be the method with the smallest relative bias and smallest relative root mean squared error at all the sample sizes, in estimating the theoretical 99th percentile.

Type 7 method, currently used in Analytica, underestimates the theoretical 99th percentile in all the scenarios tested, especially for sample sizes lower than 20,000 values.

REFERENCES

Dielmann T., Lowry C. Pfaffenberger R., A comparison of quantile estimators, Communication in Statistics – Simulation and Computation, Vol. 23 (2), pp. 355-371, 1994.

Hahn G. J., Meeker W. Q., Statistical Intervals: A Guide for Practitioners, New York, John Wiley & Sons, 1991.

Hyndman R. J., Fan T., Sample Quantiles in Statistical Packages, The American Statistician, Vol. 50, No. 4, pp. 361-365, 1996.

Kreutzmann A. K., Estimation of sample quantiles: challenges and issues in the context of income and wealth distributions, AStA Wirtsch Sozialstat Arch, No. 12, pp. 245-270, 2018.

Langford E., Quartiles in Elementary Statistics, Journal of Statistics Education, Vol. 14, No. 3, 2006.

Lumina Decision Systems Inc., User Guide Analytica 4.6, 13 May 2015, www.lumina.com.

Makkonen L., Pajari M., Defining Sample Quantiles by the True Rank Probability, Journal of Probability and Statistics, Vol. 2014, Article ID 326579, 2014.

Oracle, Crystal Ball 11.1.2 User Guide, 2010, www.oracle.com.

Palisade, @Risk 7.0 User Guide, July 2016, www.palisade.com.

Parrish R. S., Comparison of quantile estimators in normal sampling, Biometrics, Vol. 46, No. 1, pp 247-257, 1990.

Schoonjans F., DeBacquer D., Schmid P., Estimation of Population Percentiles, Epidemiology, Vol. 22, No. 5, pp 750-751, 2011.

Stuart A., Ord K., Kendall's Advanced Theory of Statistics, London: Arnold, ISBN 0340614307, 1994.

Vose Software, Model Risk 5.0 Help, 2012, www.vosesoftware.com.

APPENDIX A

This appendix shows the 99th percentiles computed using Type 1 (Nearest-Rank method), Type 2 (default method in SAS), Type 4 (used in @Risk), Type 6 (used by Excel PERCENTILE.EXC function), and Type 7 (used in Analytica), and the Vose software method, for the following 10 distributions:

Uniform (1,6),

Triangular (1,3,5),

Triangular (1,5,5),

Weibull (2,1),

Weibull (1,10),

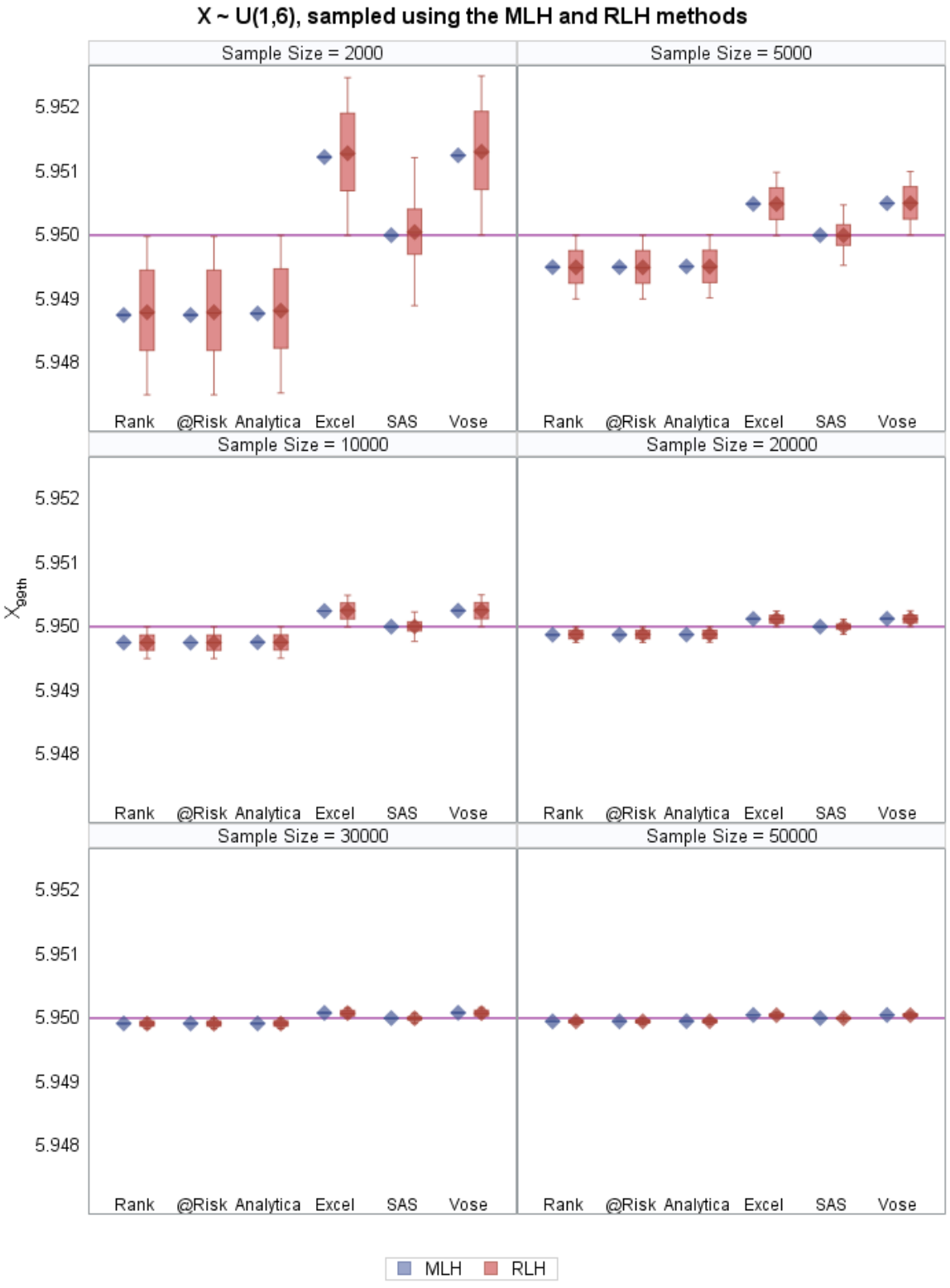
Normal (5,1),

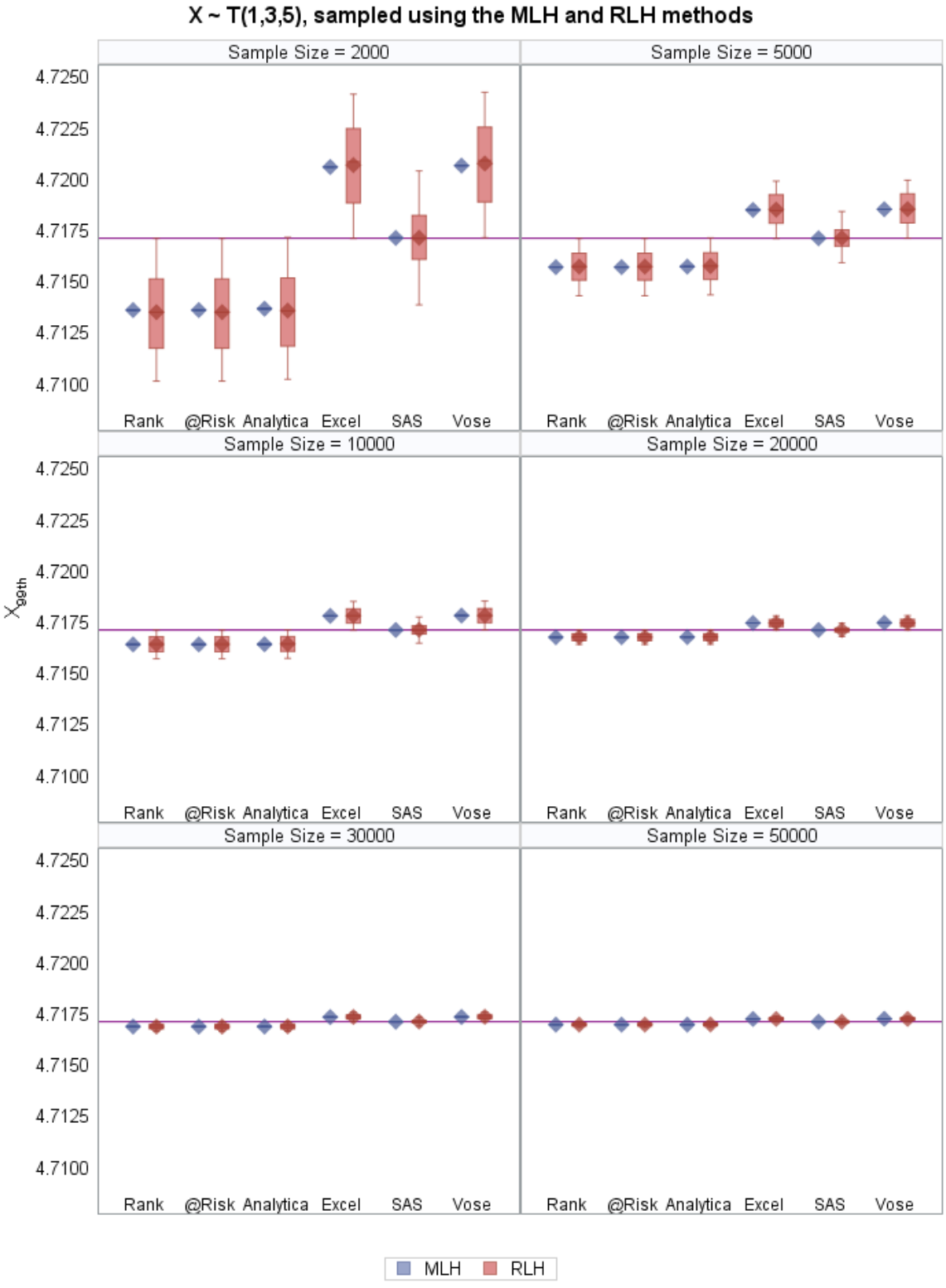
Normal (25,5),

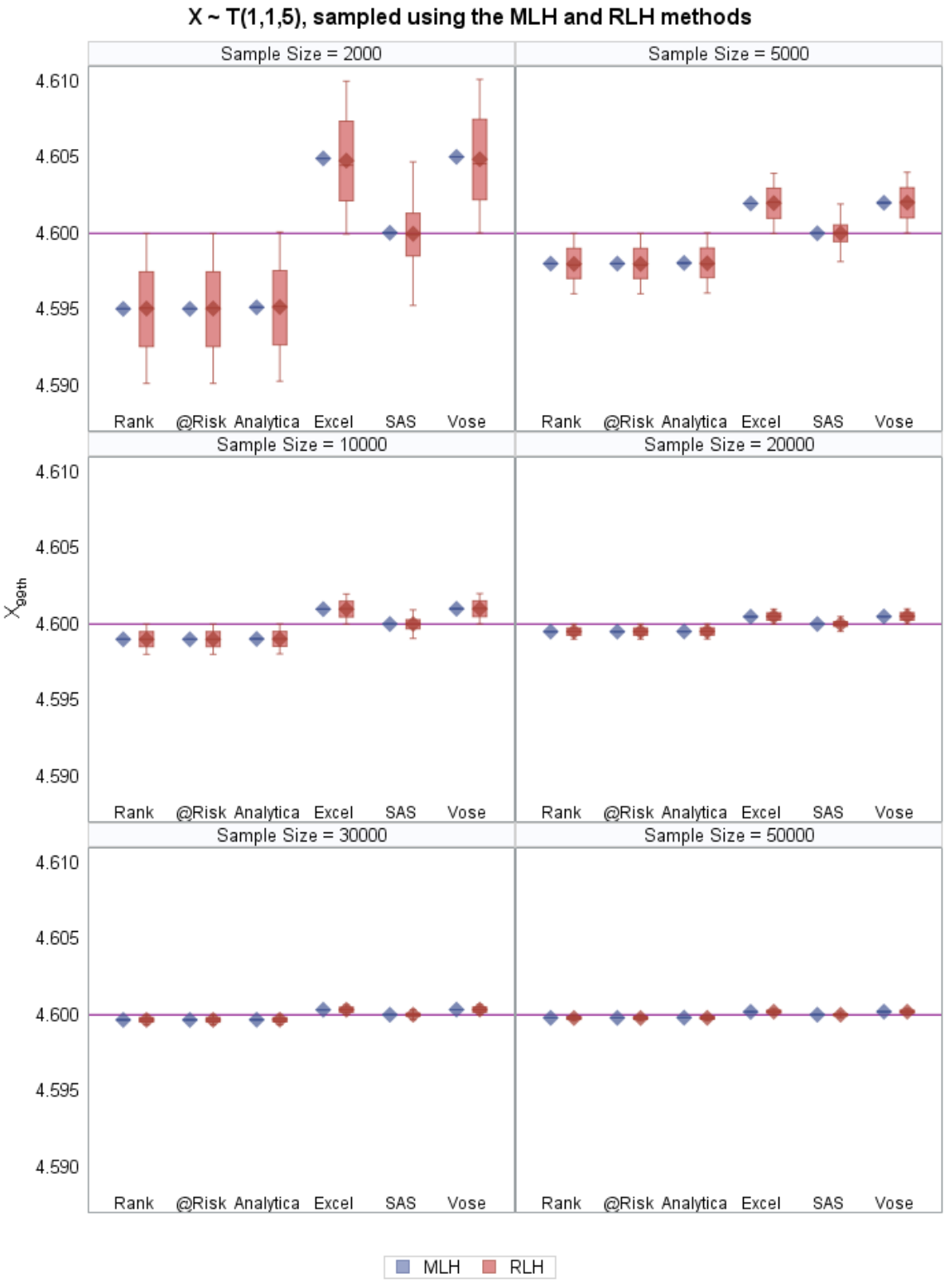
Lognormal (1,3),

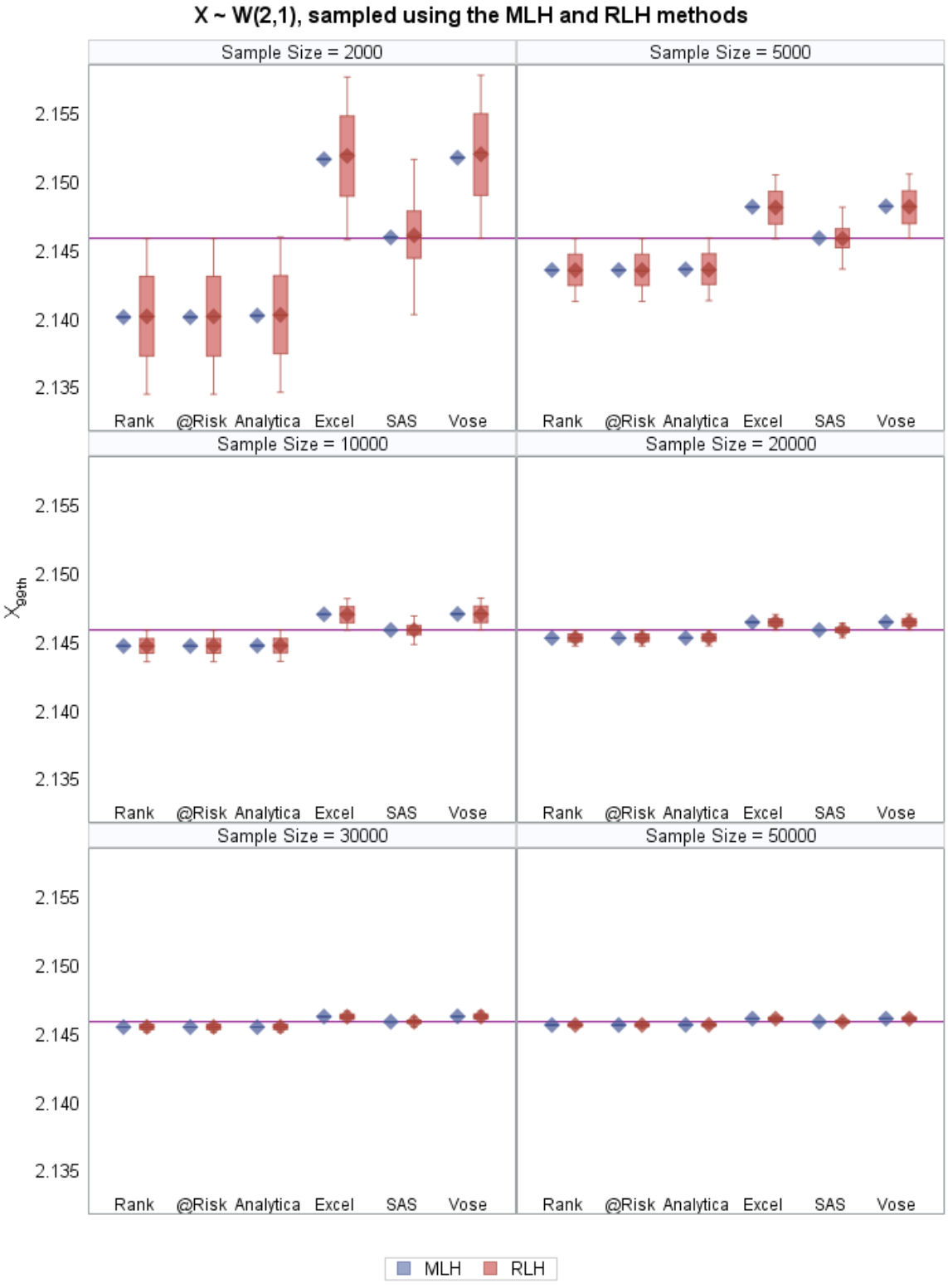
Lognormal (3,6),

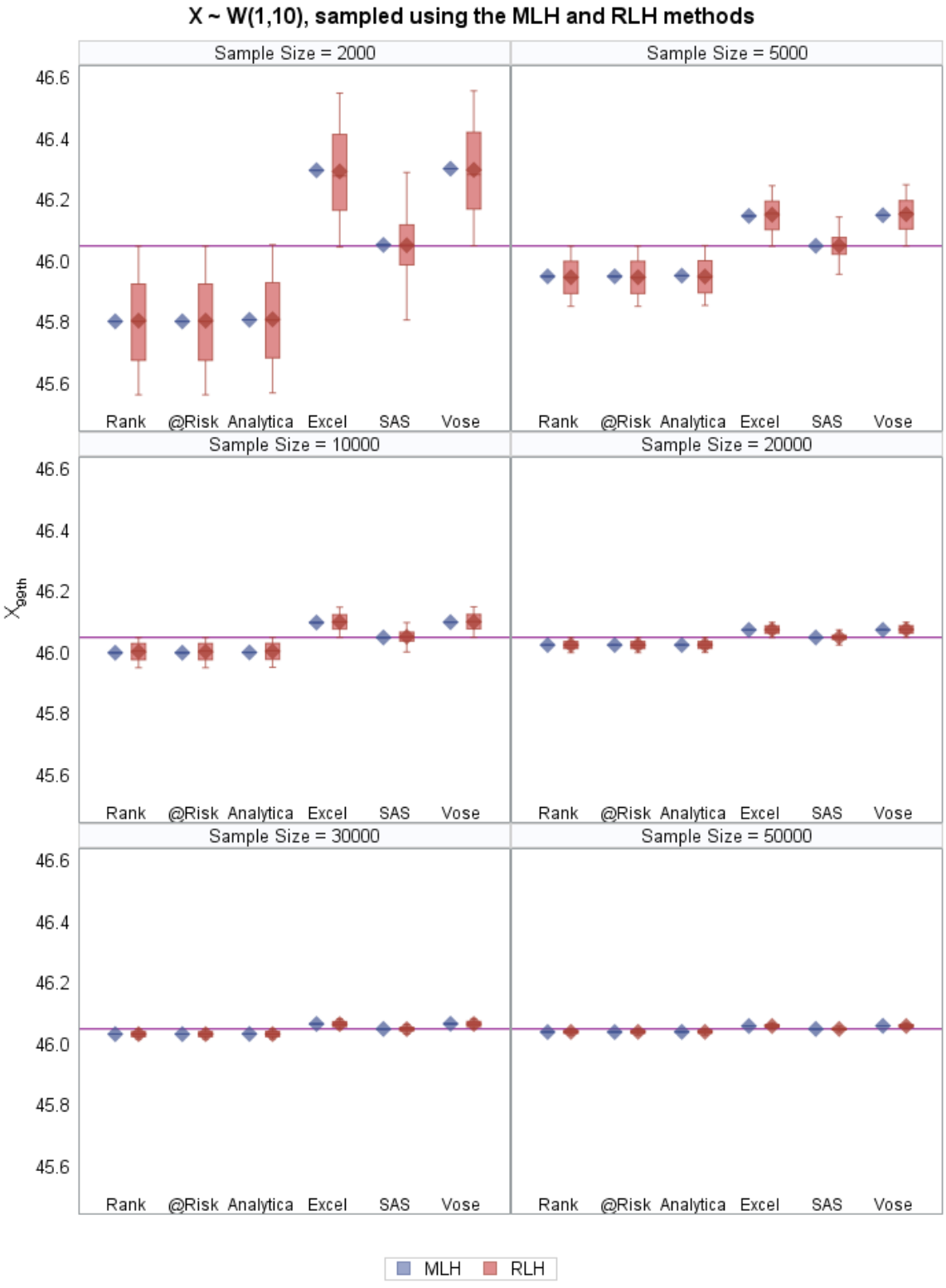
Lognormal (3,10).

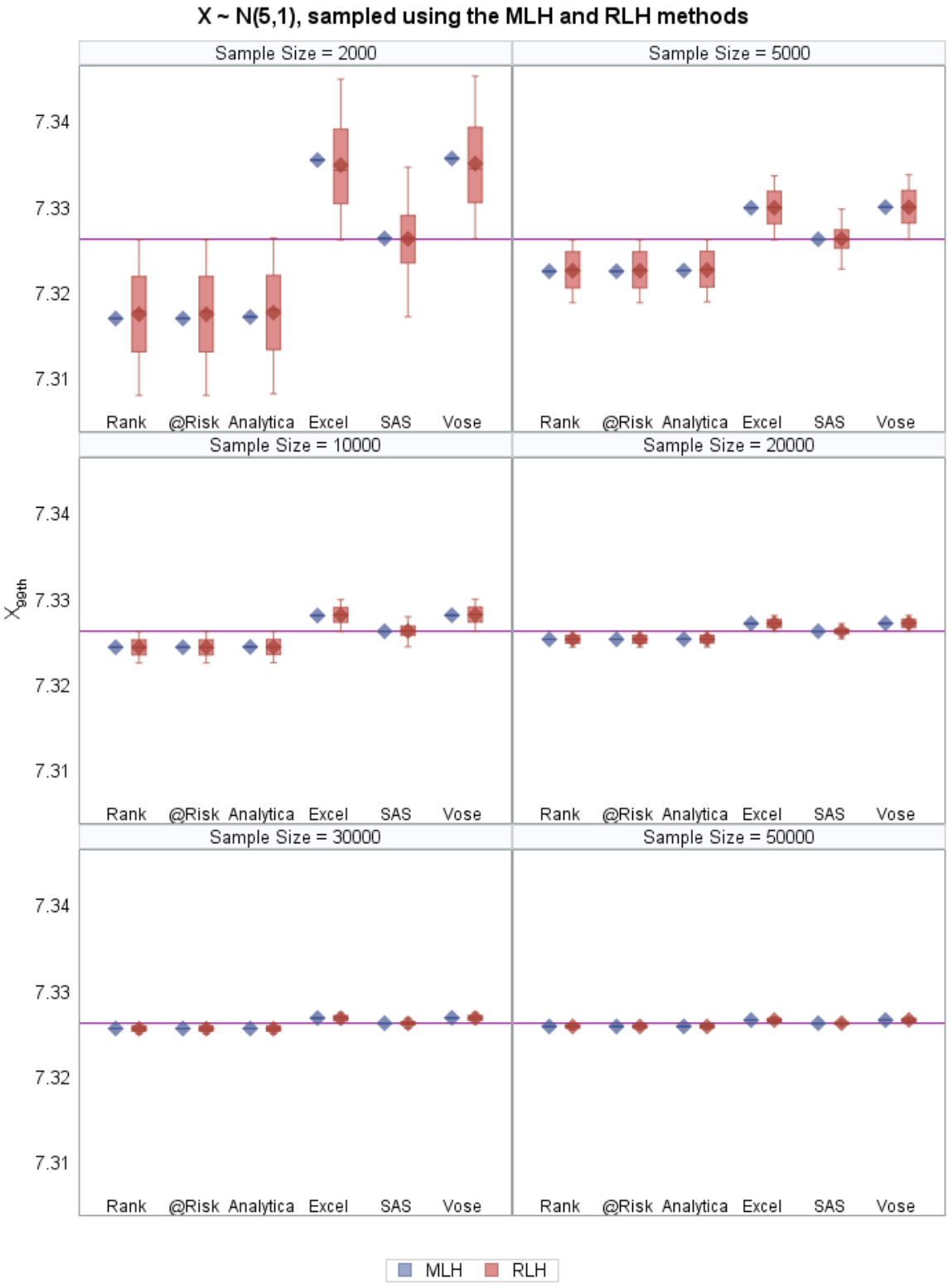


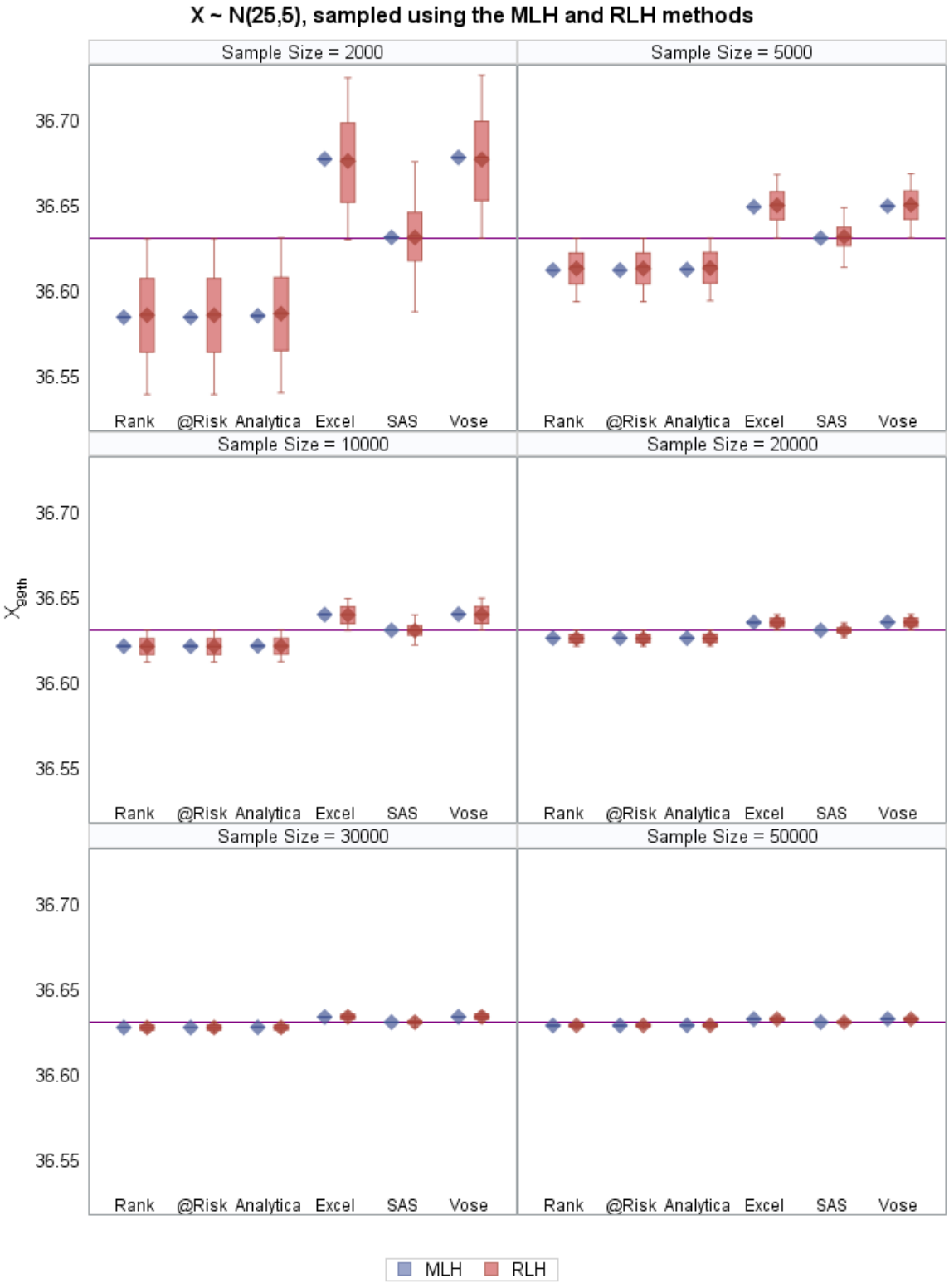


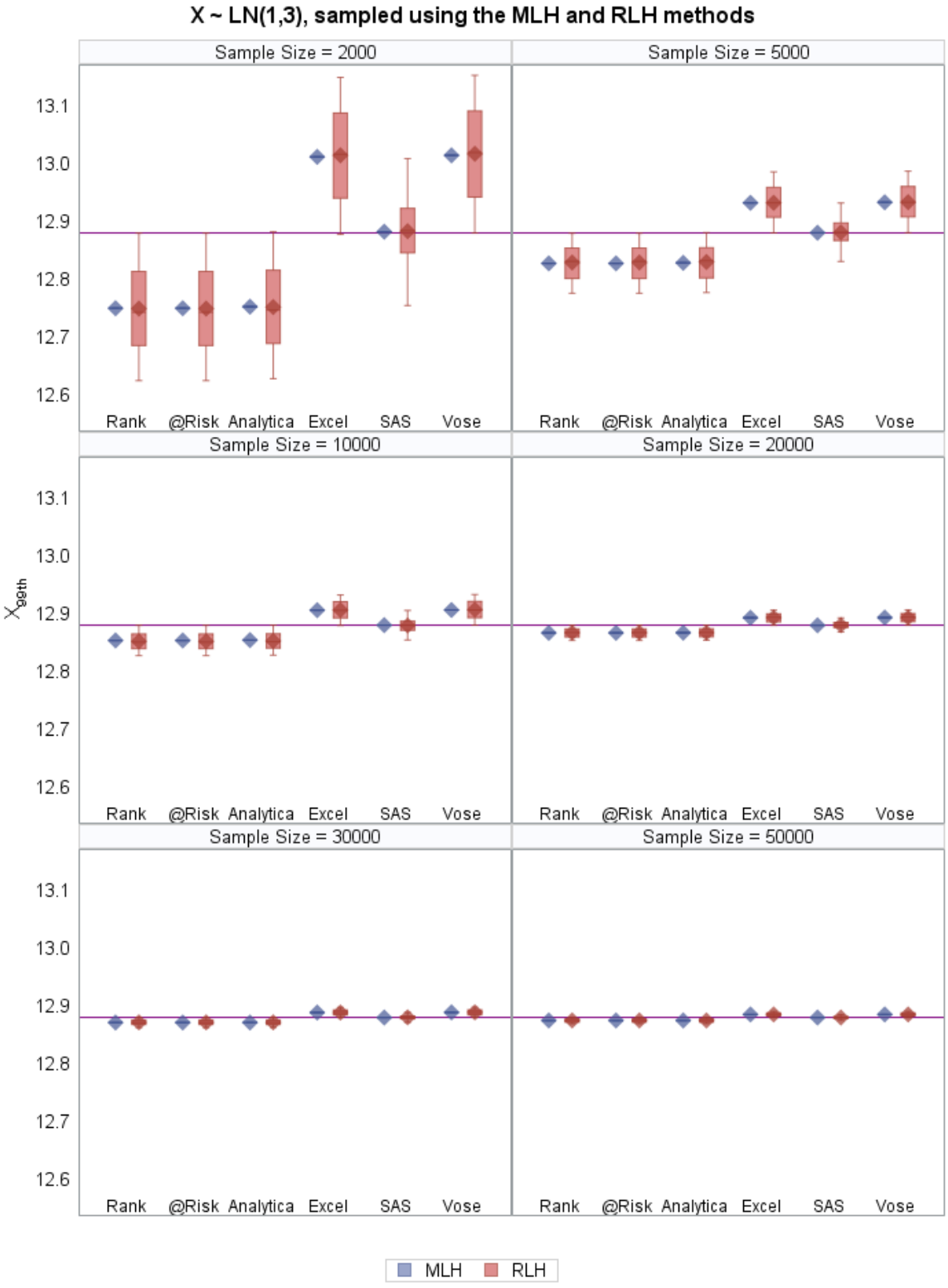


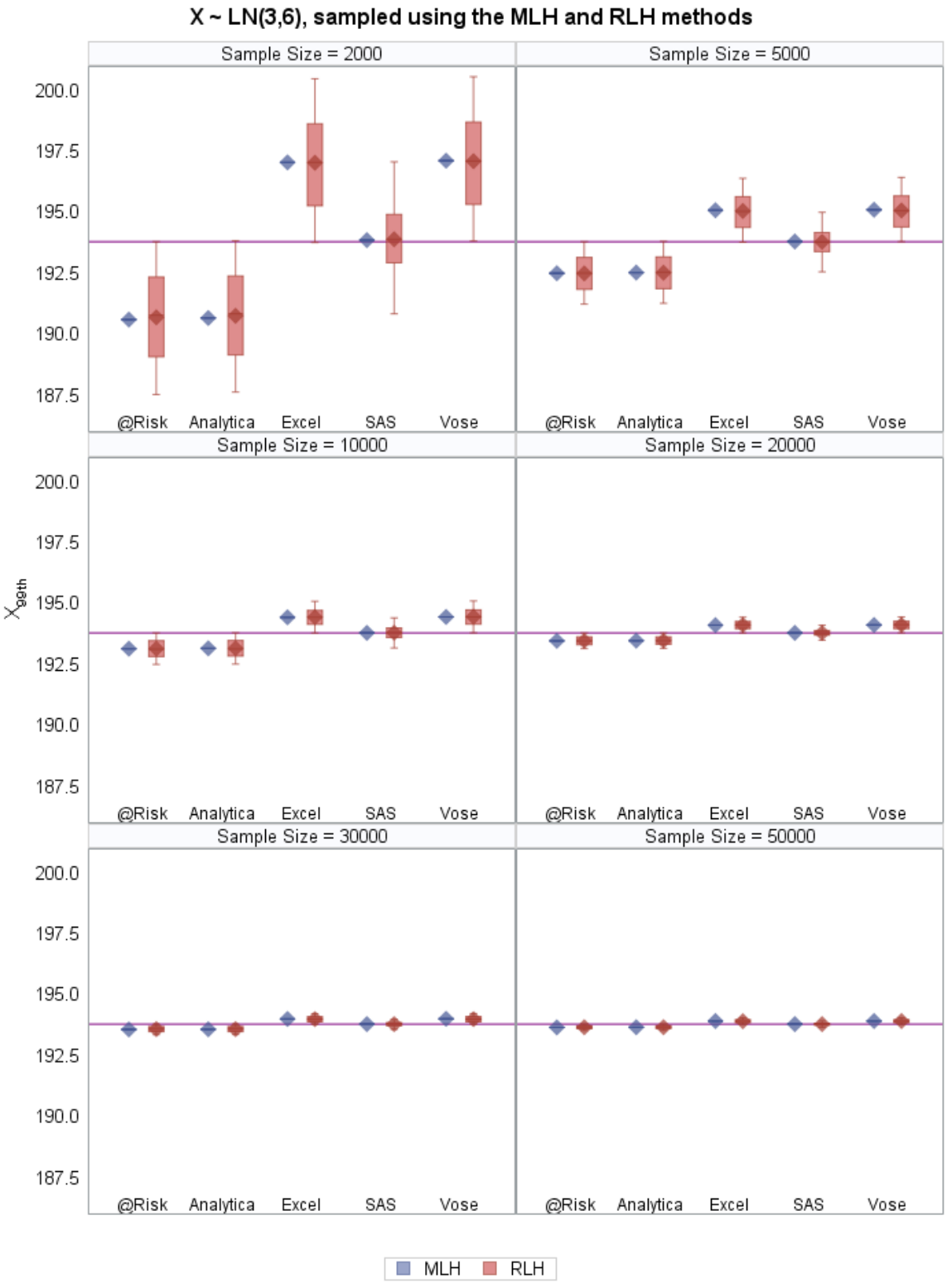


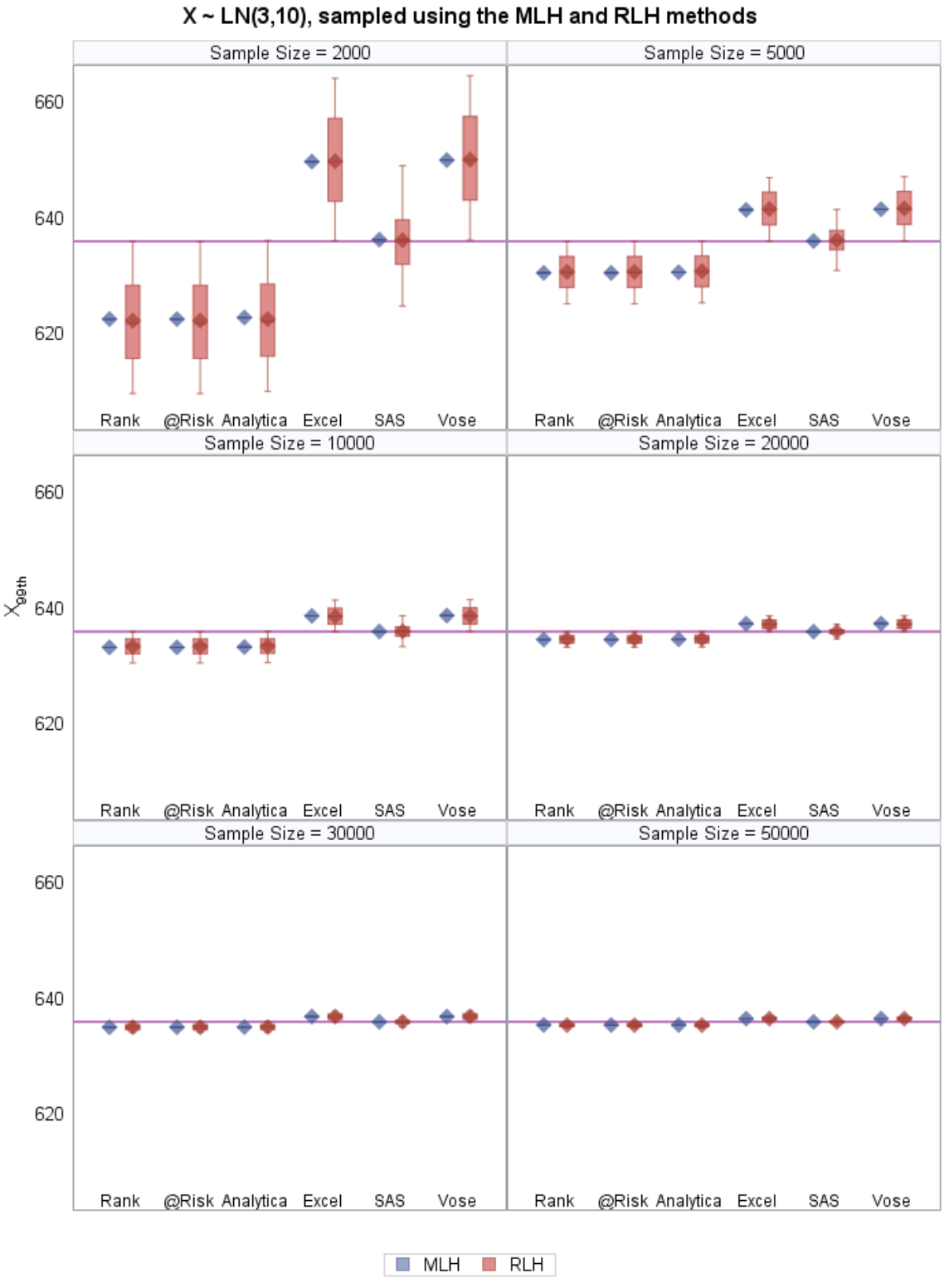












APPENDIX B

This appendix shows the 99th percentiles computed using Type 2 (default method in SAS) and Type 7 (used in Analytica) methods, for the sum and product of the following combinations of two distributions:

X1 + X2, X1 * X2, where X1 ~ N(5,1), and X2 ~ N(5,1),

X1 + X2, X1 * X2, where X1 ~ N(25,5), and X2 ~ N(25,5),

X1 + X2, X1 * X2, where X1 ~ N(5,1), and X2 ~ N(25,5),

X1 + X2, X1 * X2, where X1 ~ LN(1,3), and X2 ~ LN(1,3),

X1 + X2, X1 * X2, where X1 ~ LN(3,6), and X2 ~ LN(3,6),

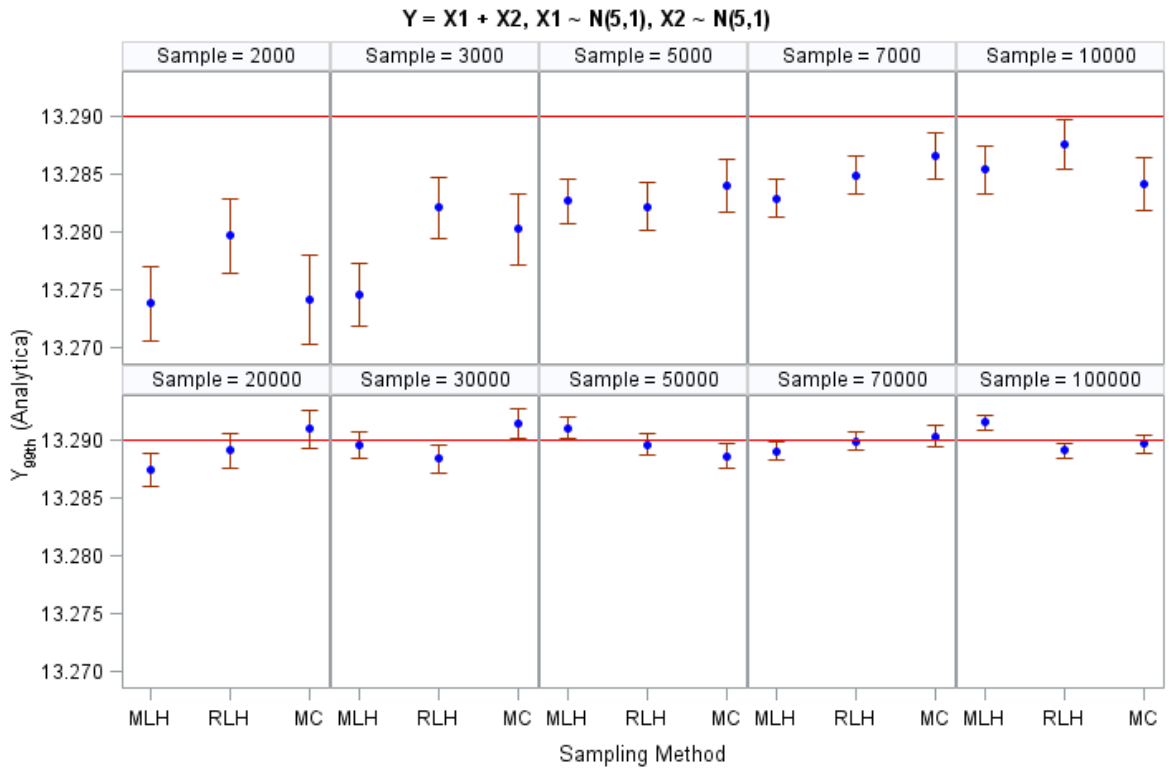
X1 + X2, X1 * X2, where X1 ~ LN(1,3), and X2 ~ LN(3,6),

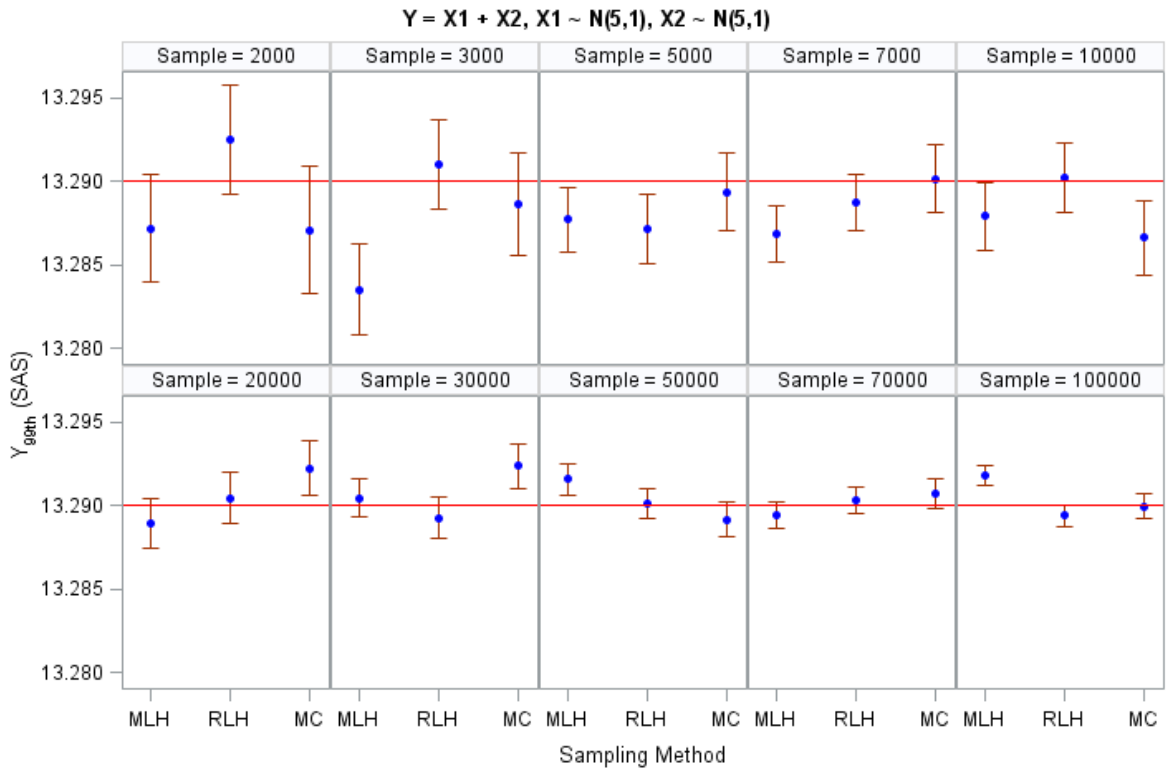
X1 + X2, X1 * X2, where X1 ~ N(5,1), and X2 ~ LN(1,3),

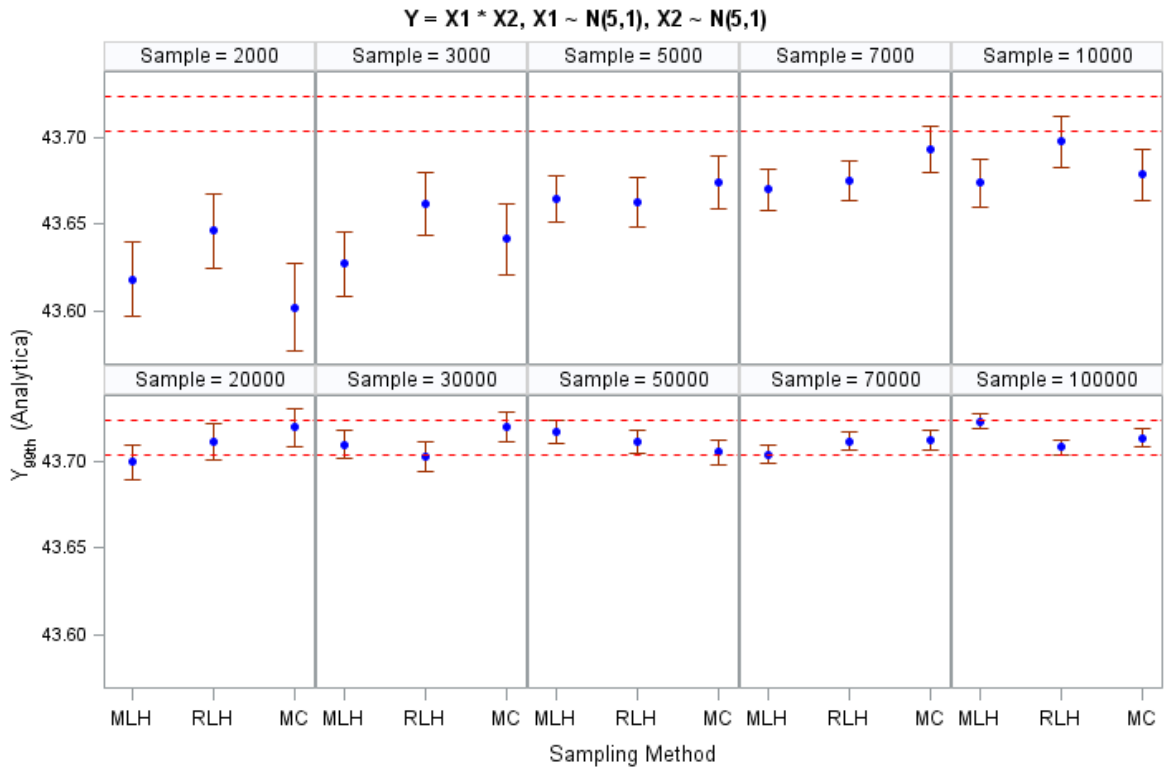
X1 + X2, X1 * X2, where X1 ~ N(5,1), and X2 ~ LN(3,6),

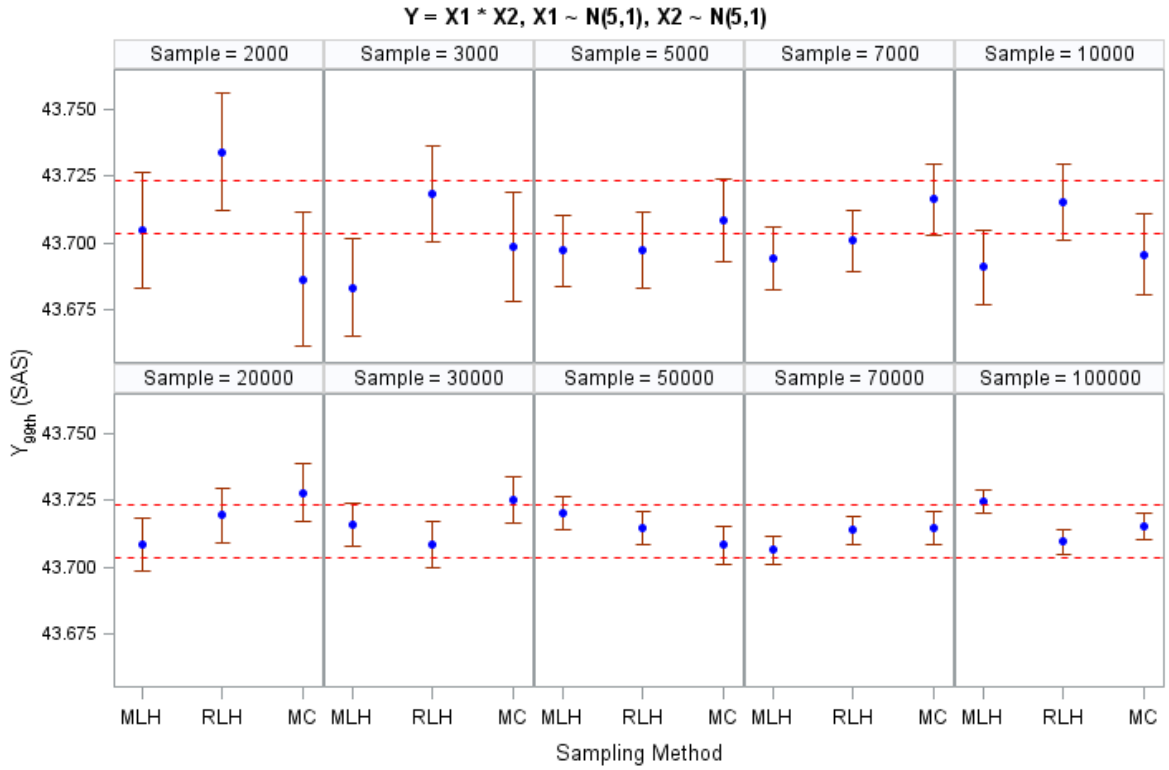
X1 + X2, X1 * X2, where X1 ~ N(25,5), and X2 ~ LN(1,3),

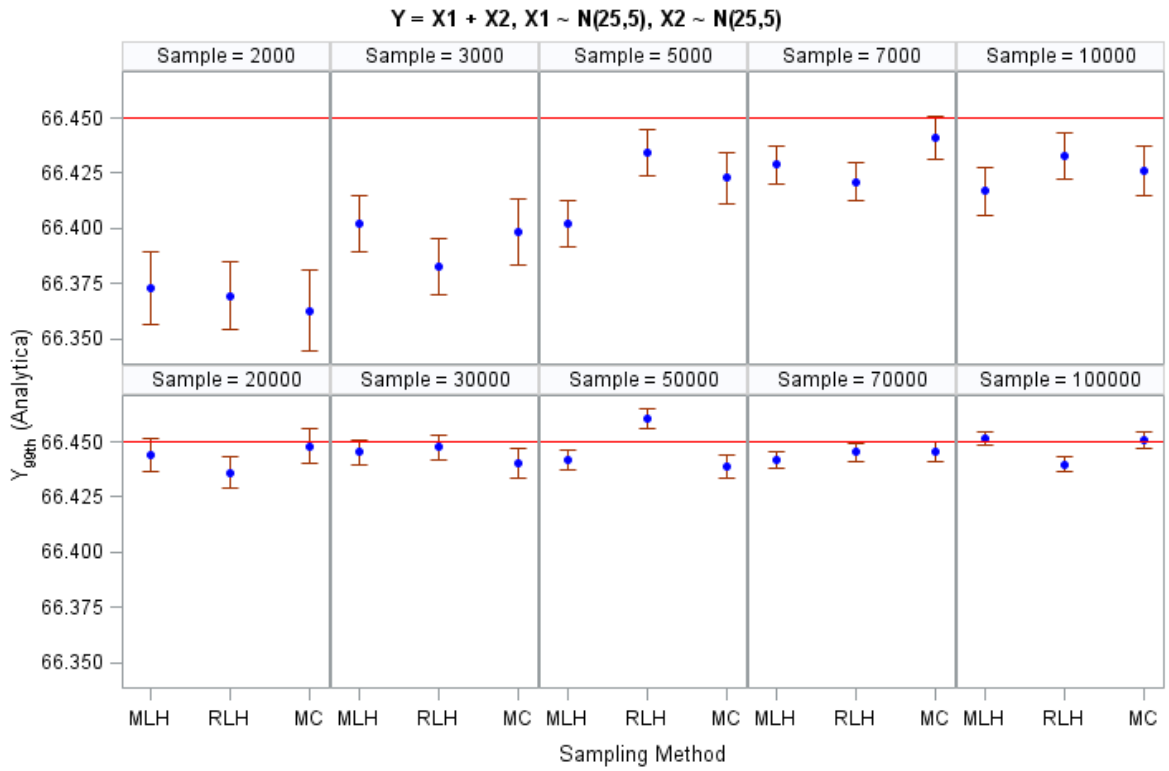
X1 + X2, X1 * X2, where X1 ~ N(25,5), and X2 ~ LN(3,6).

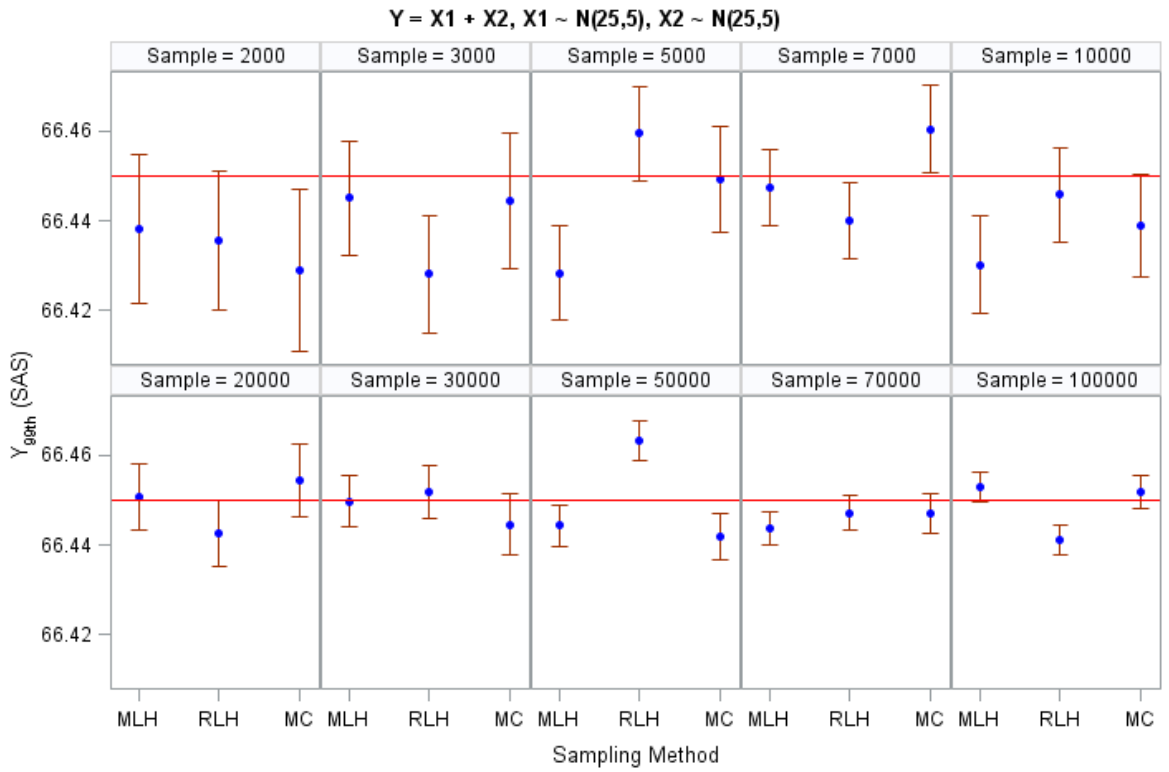


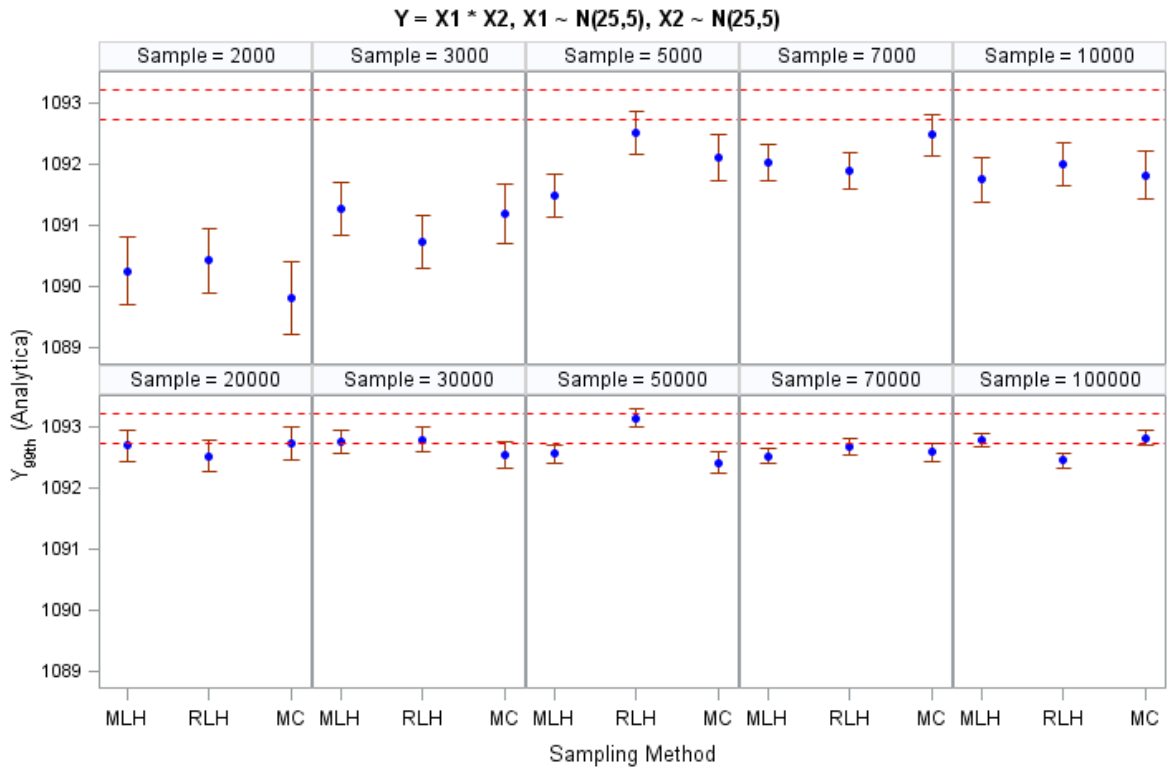


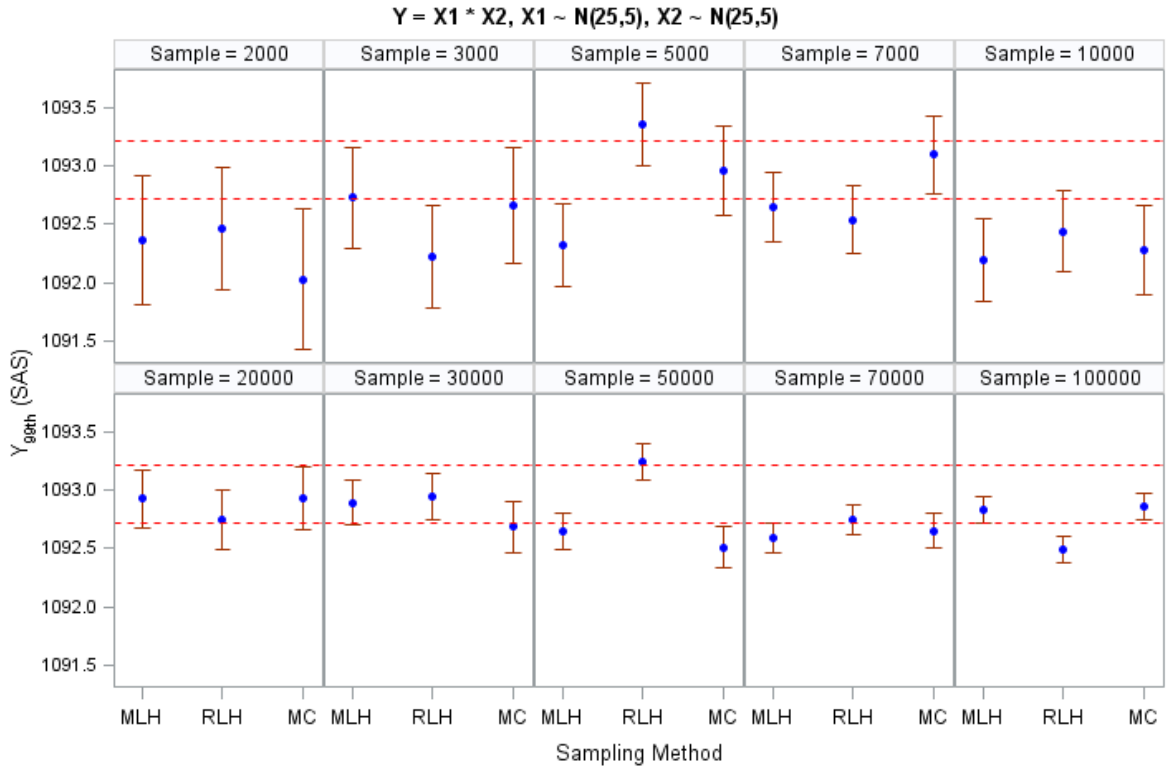


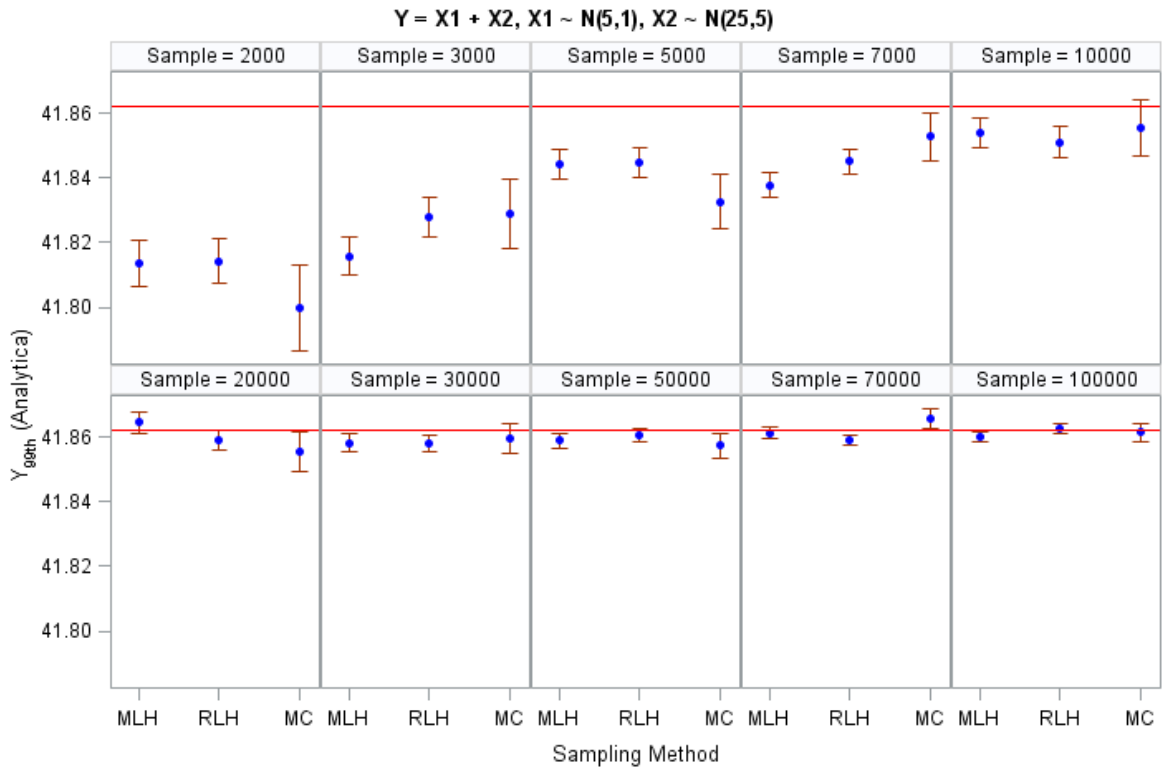


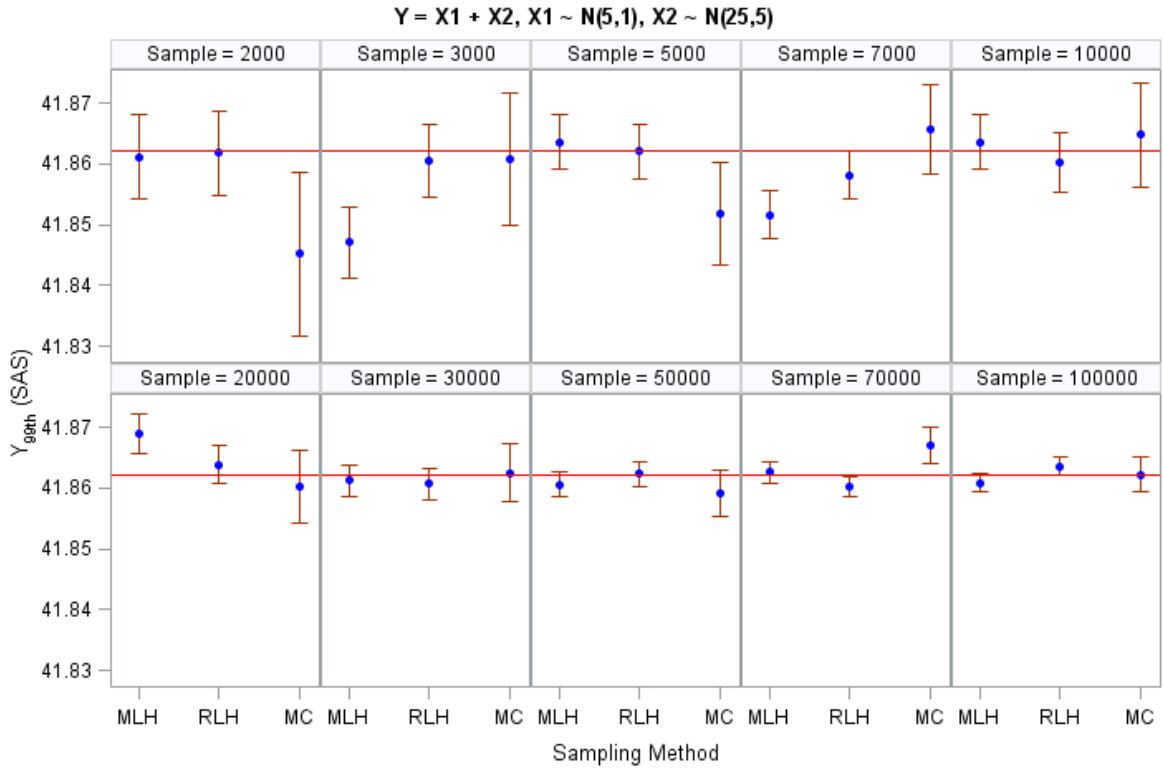


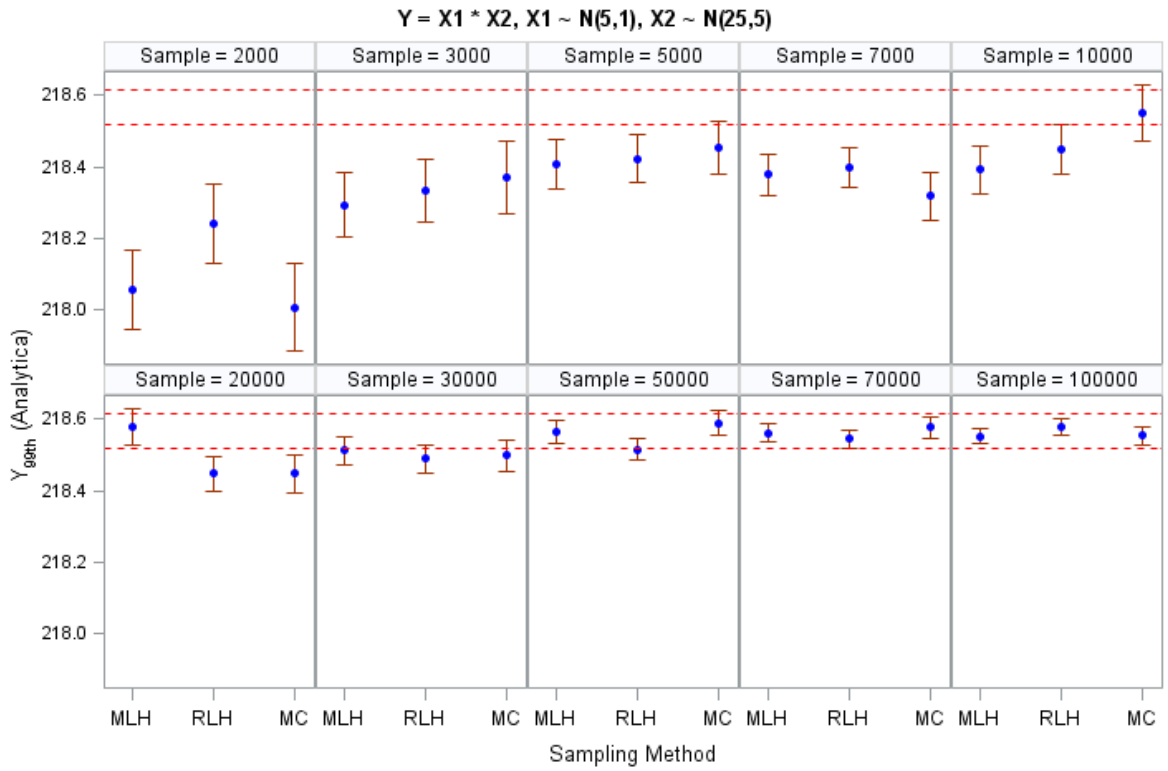


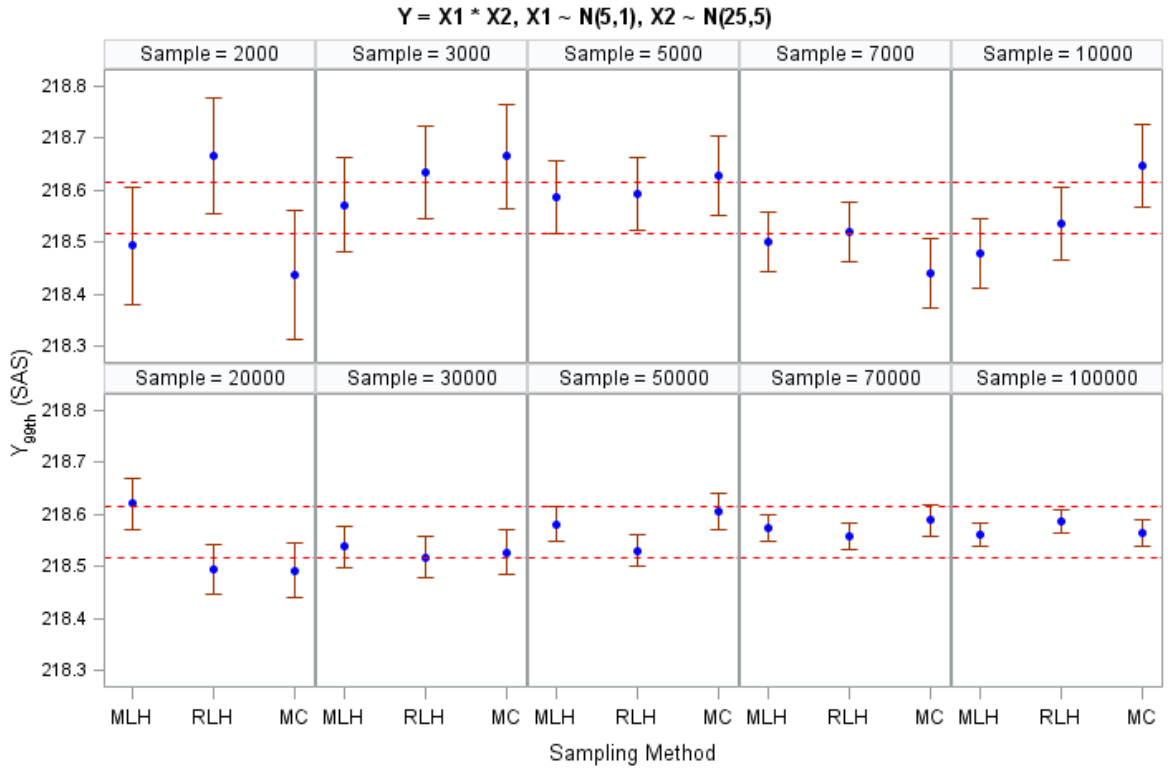


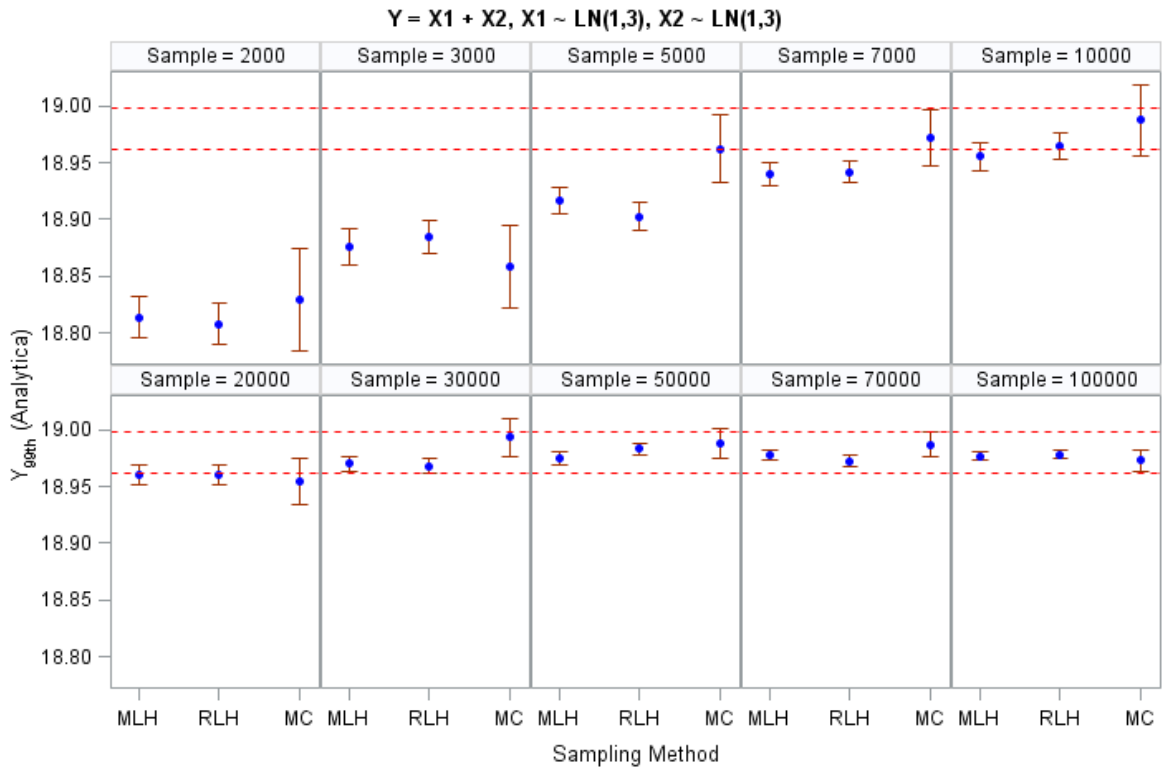


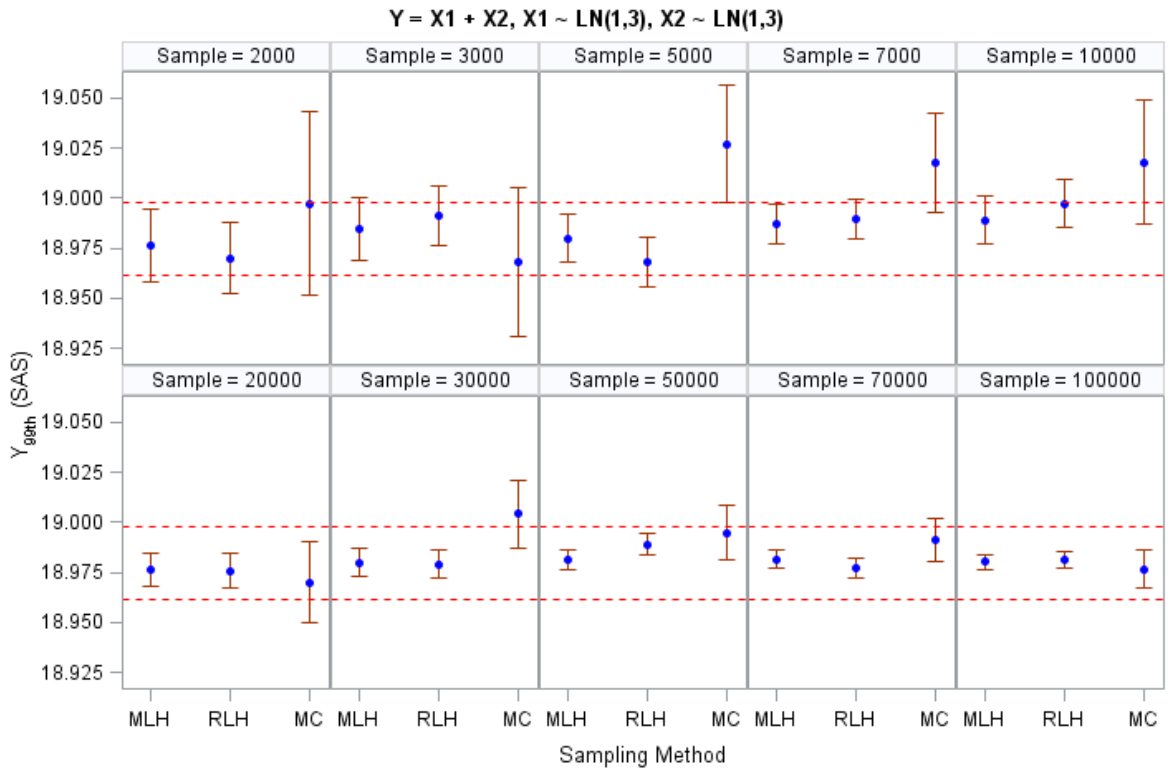


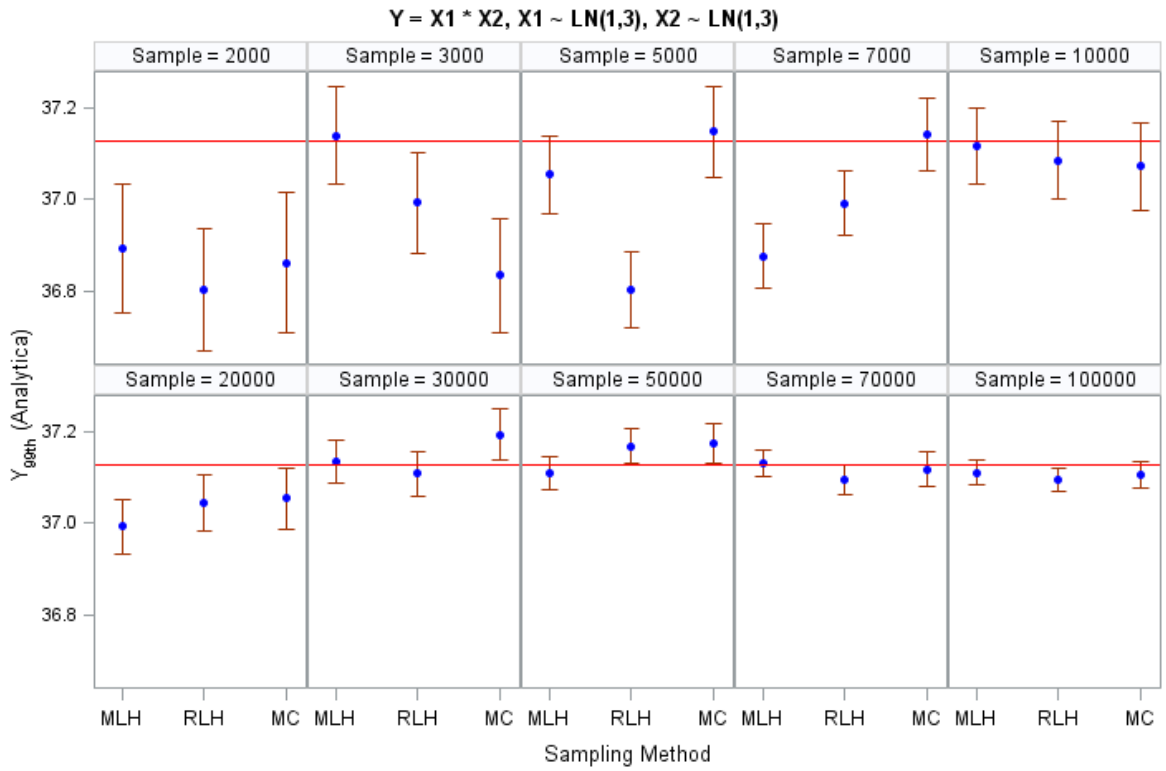


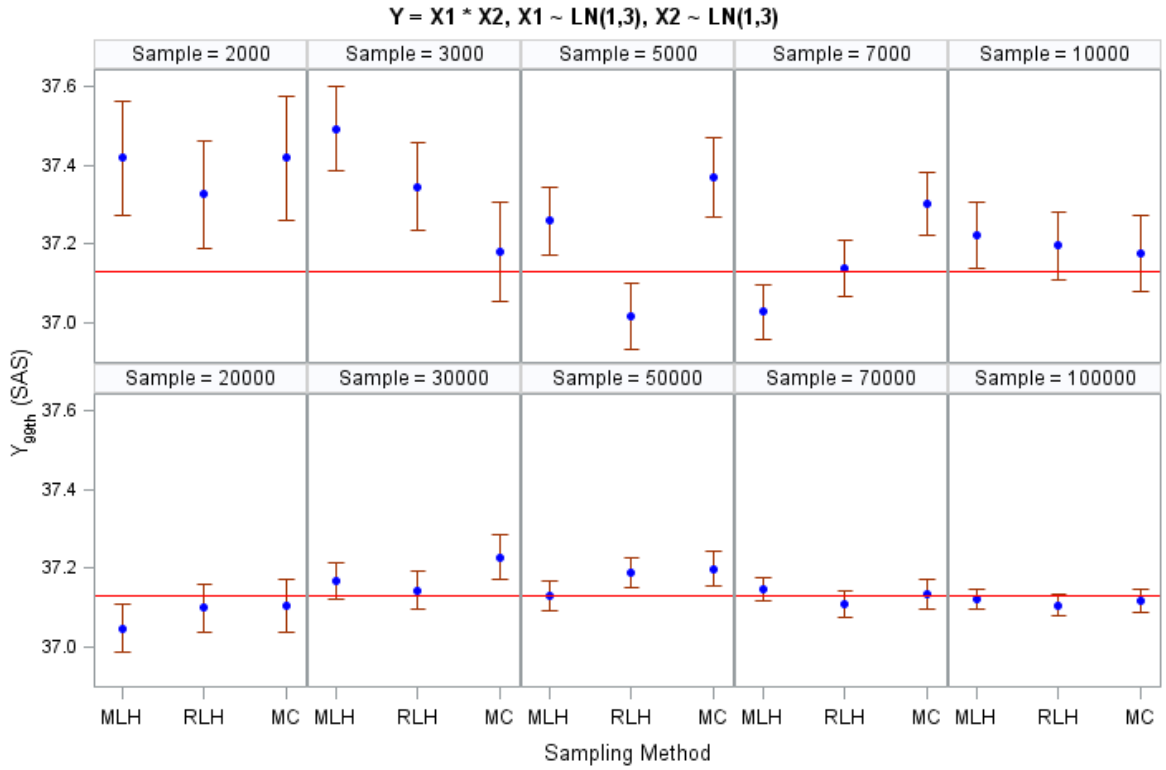


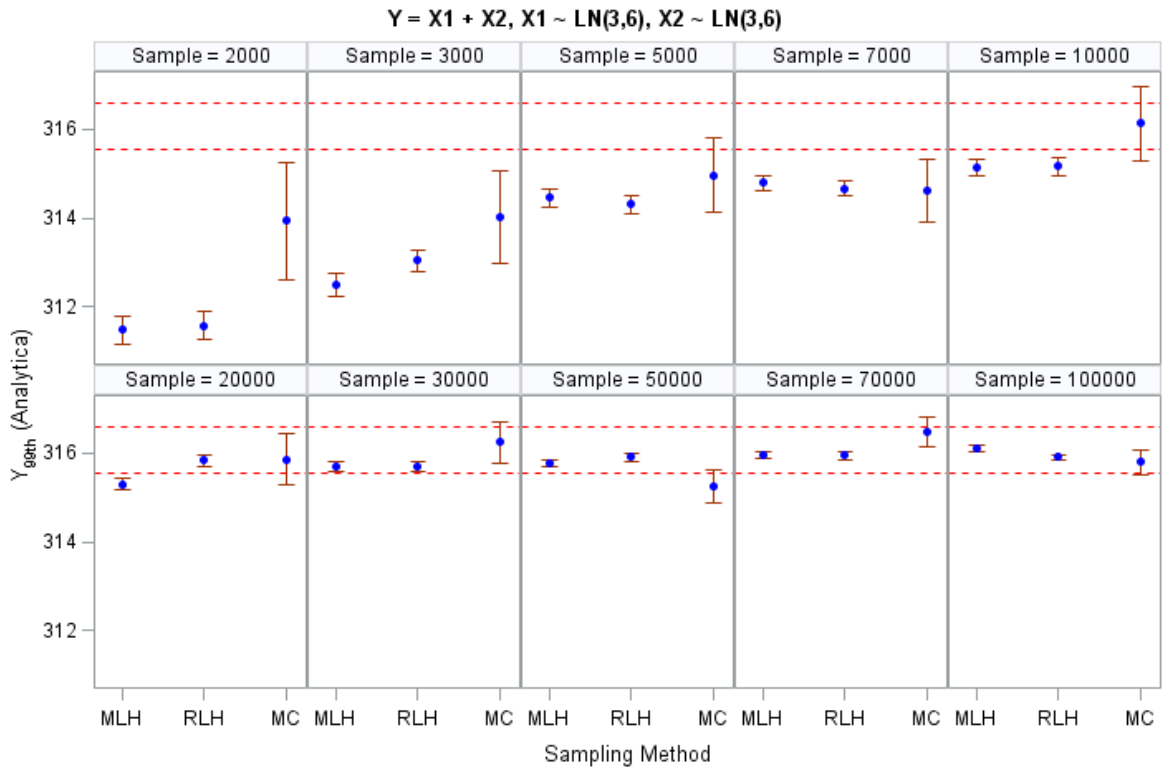


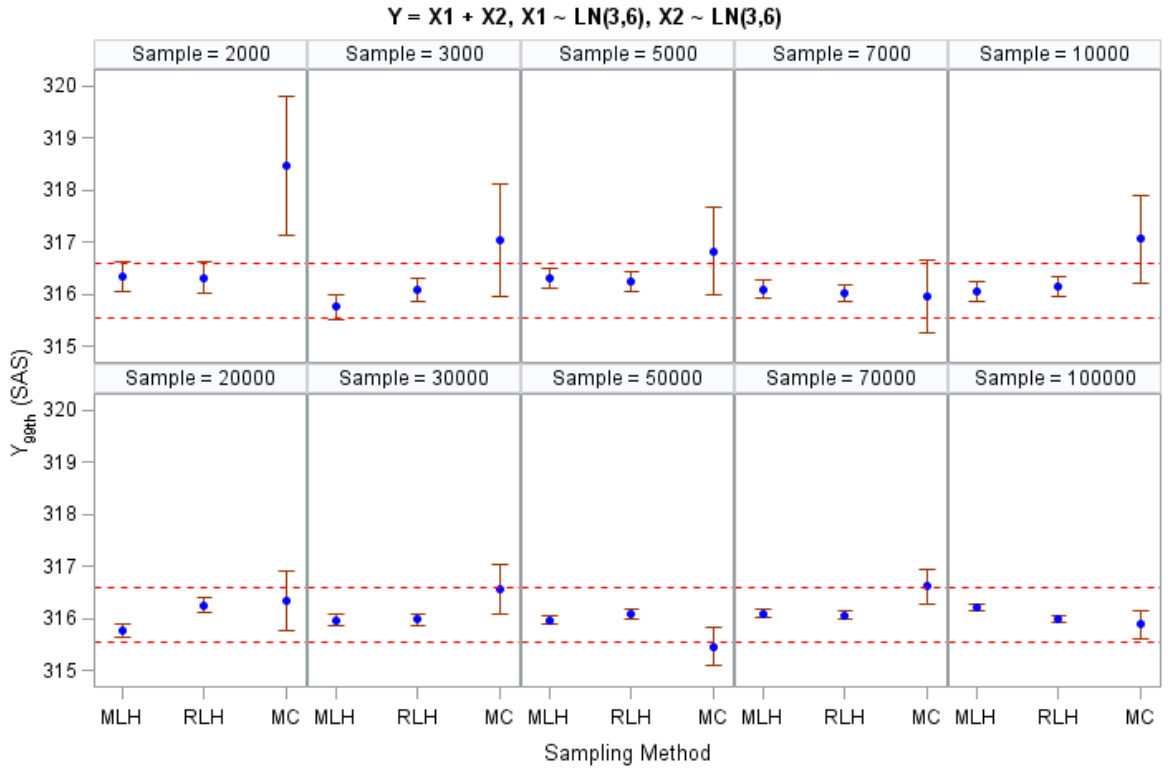


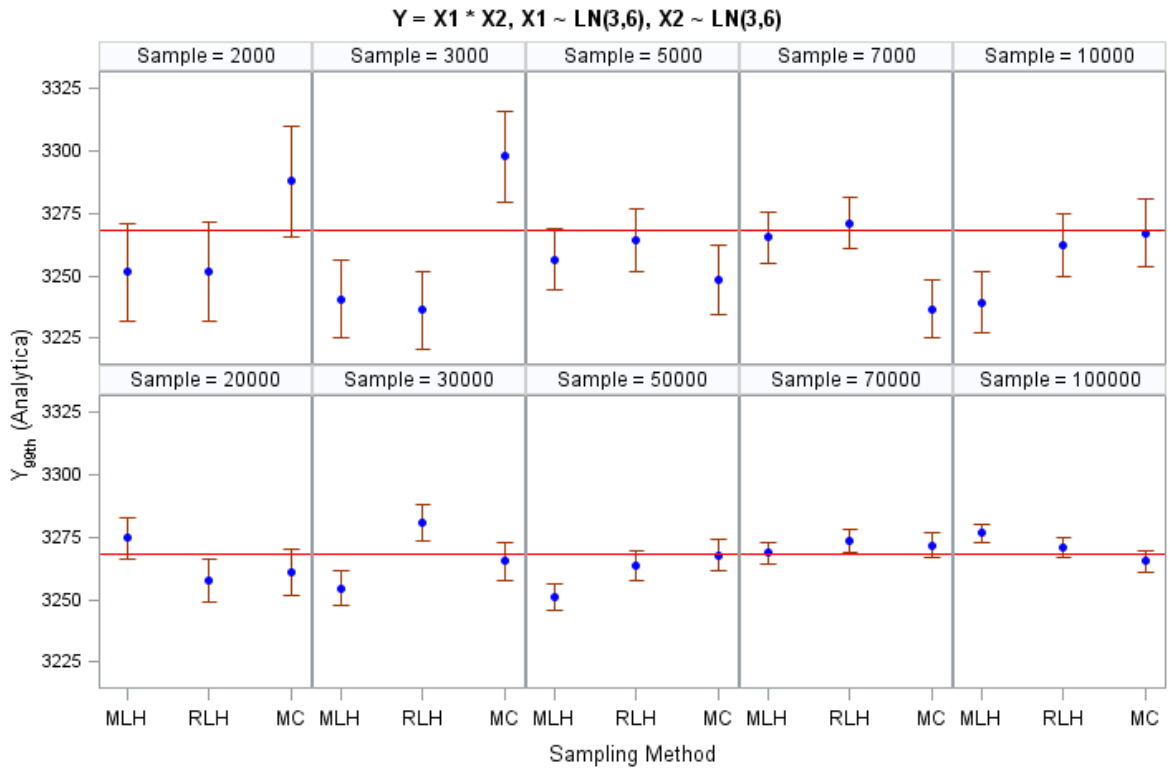


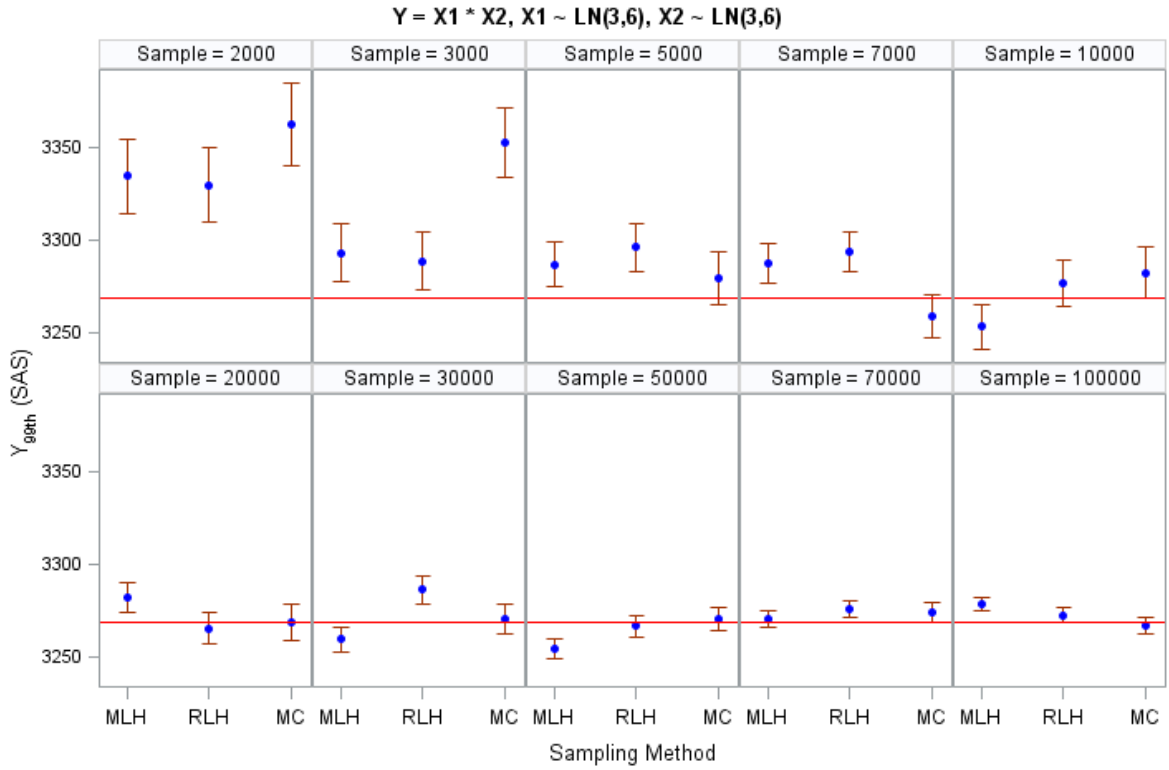


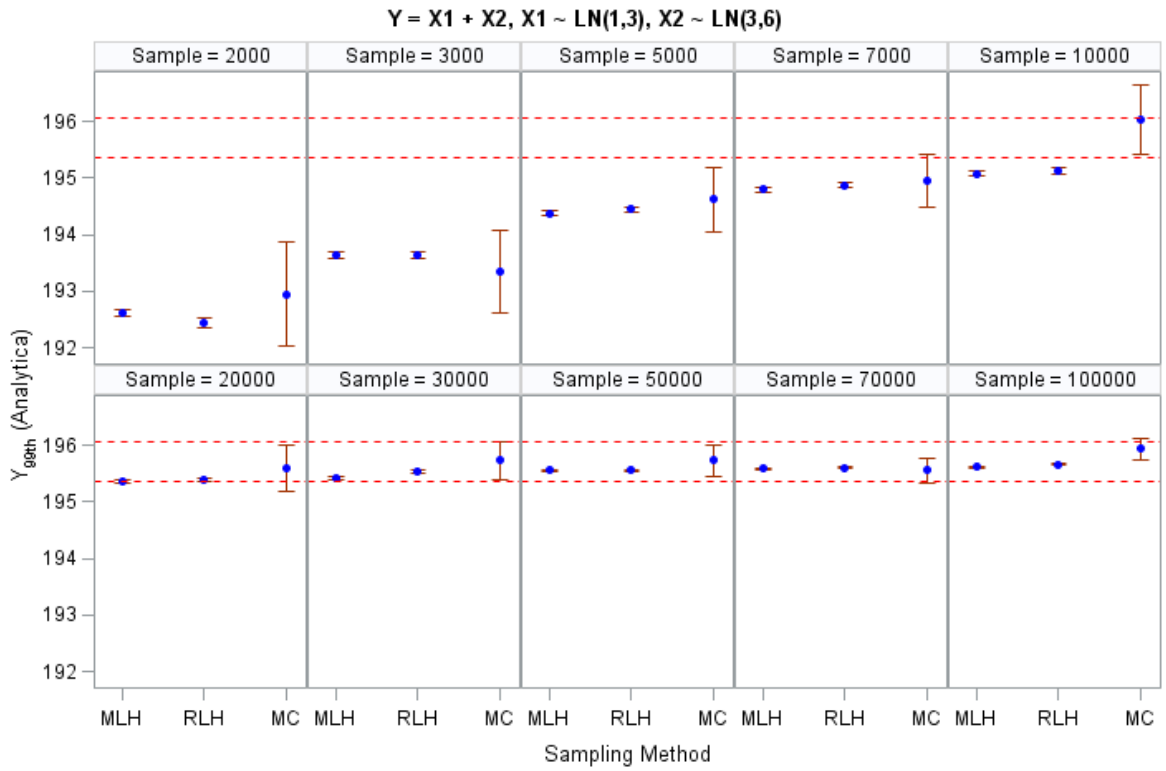


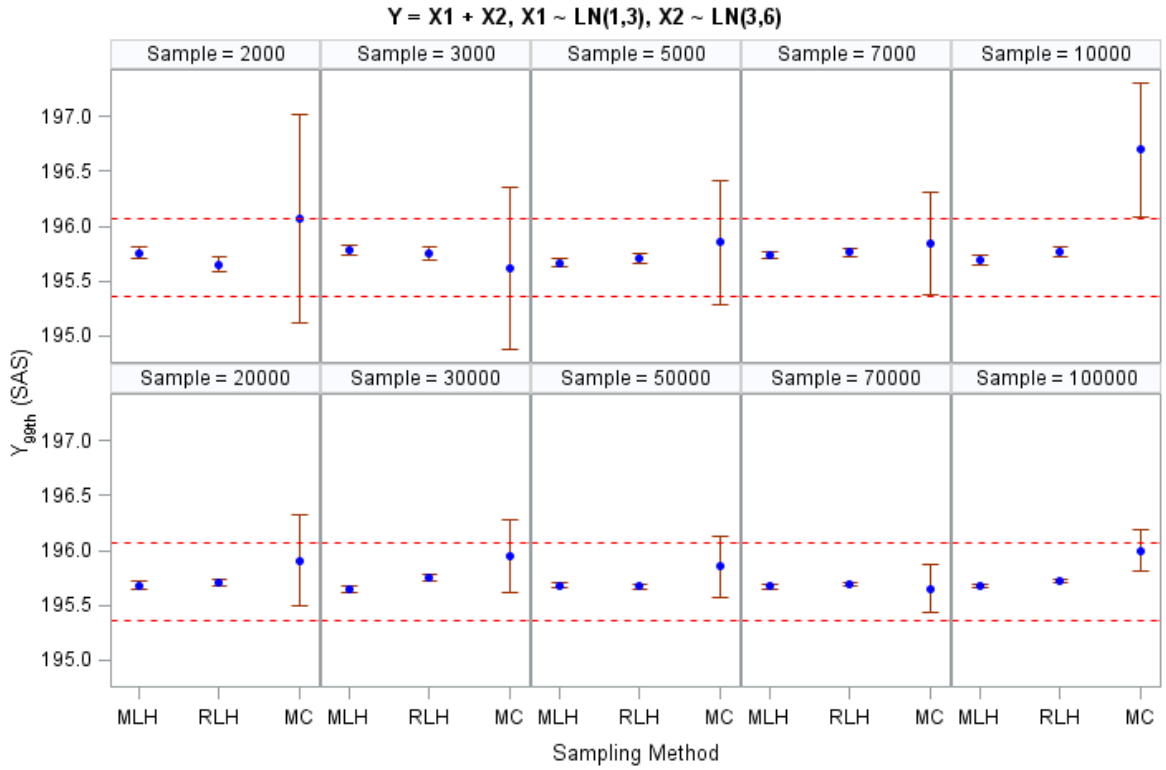


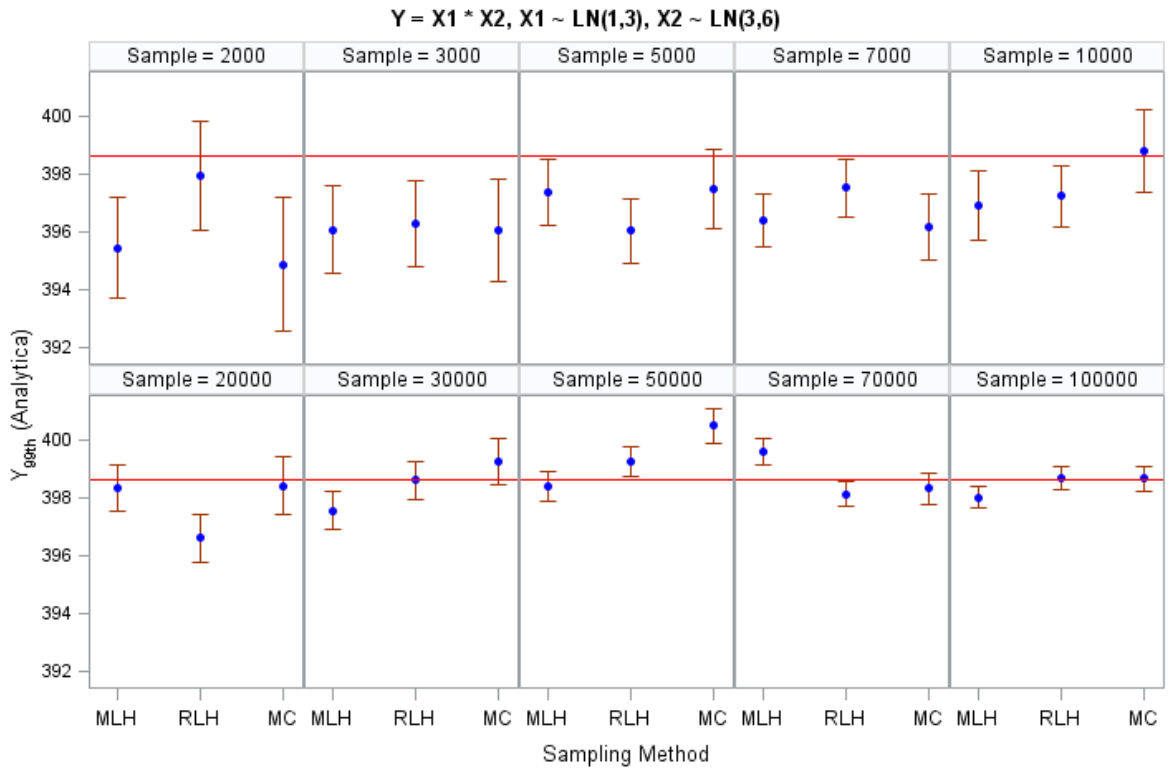


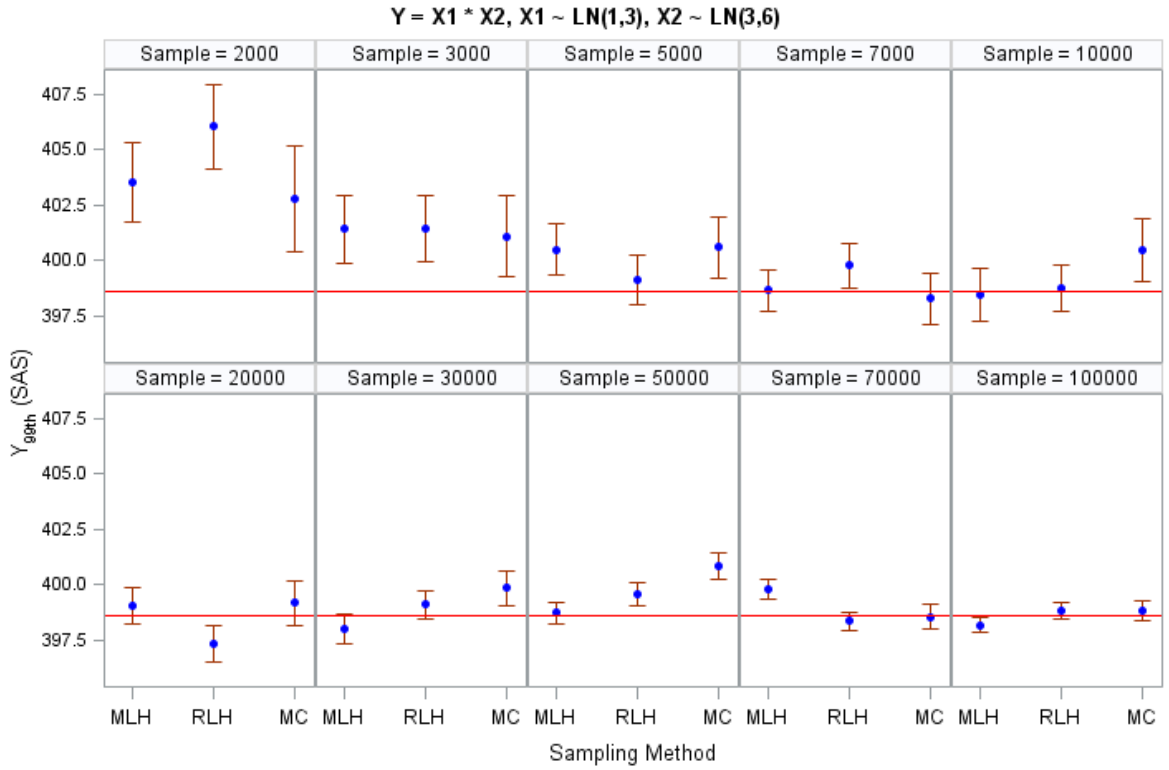


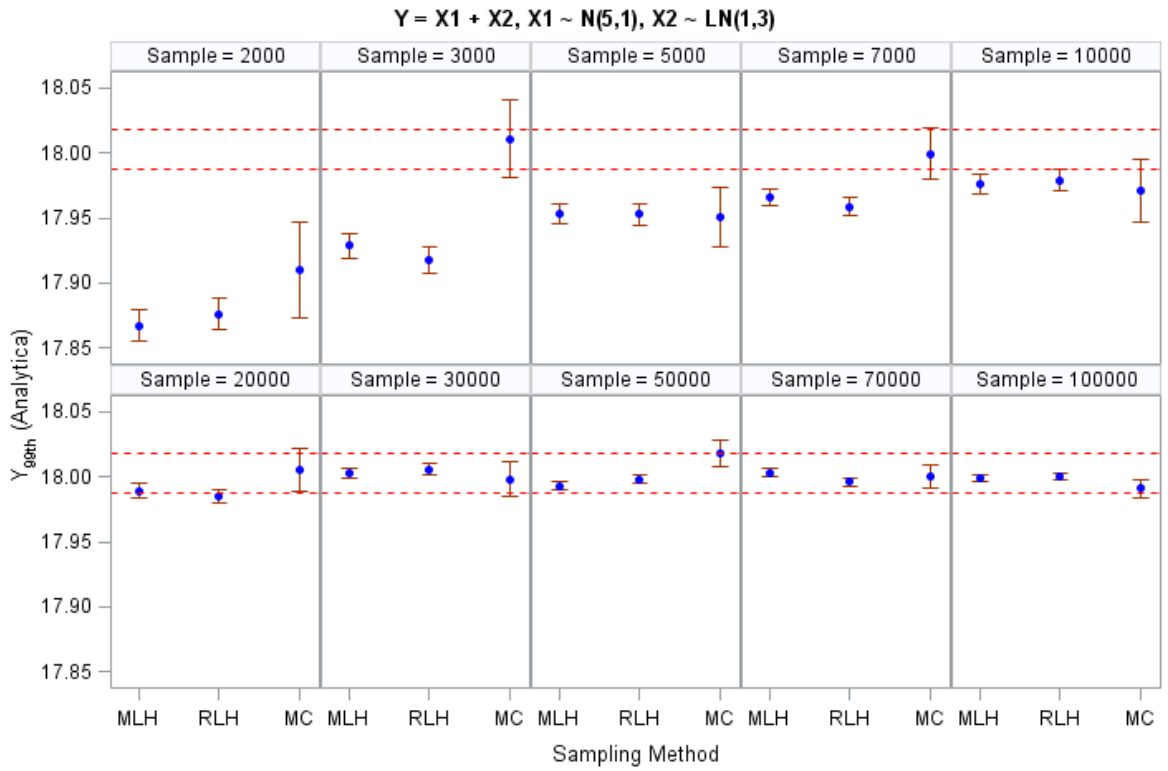


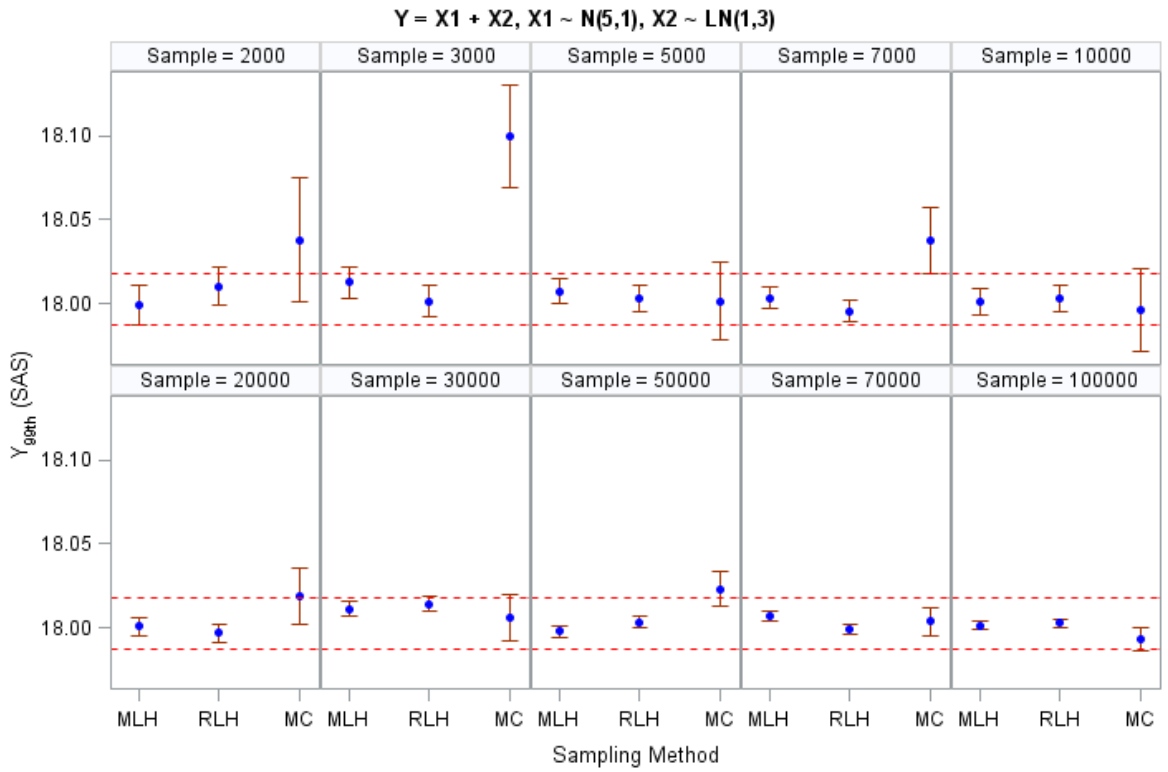


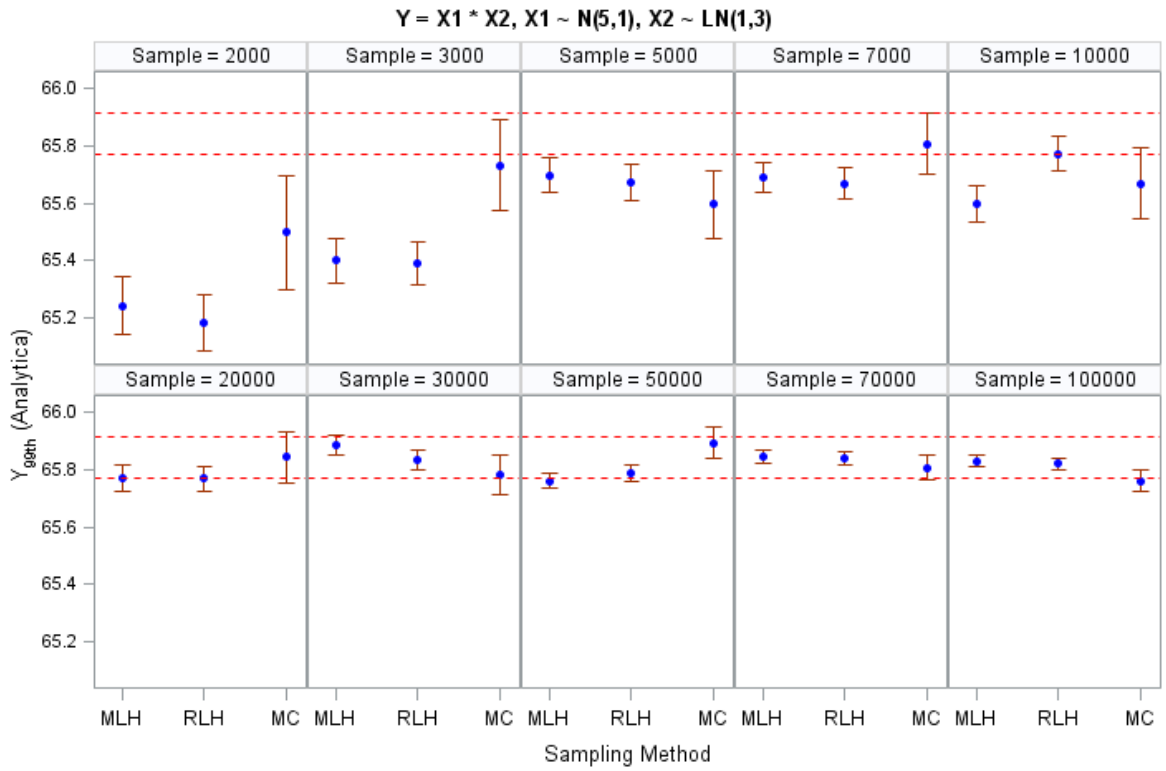


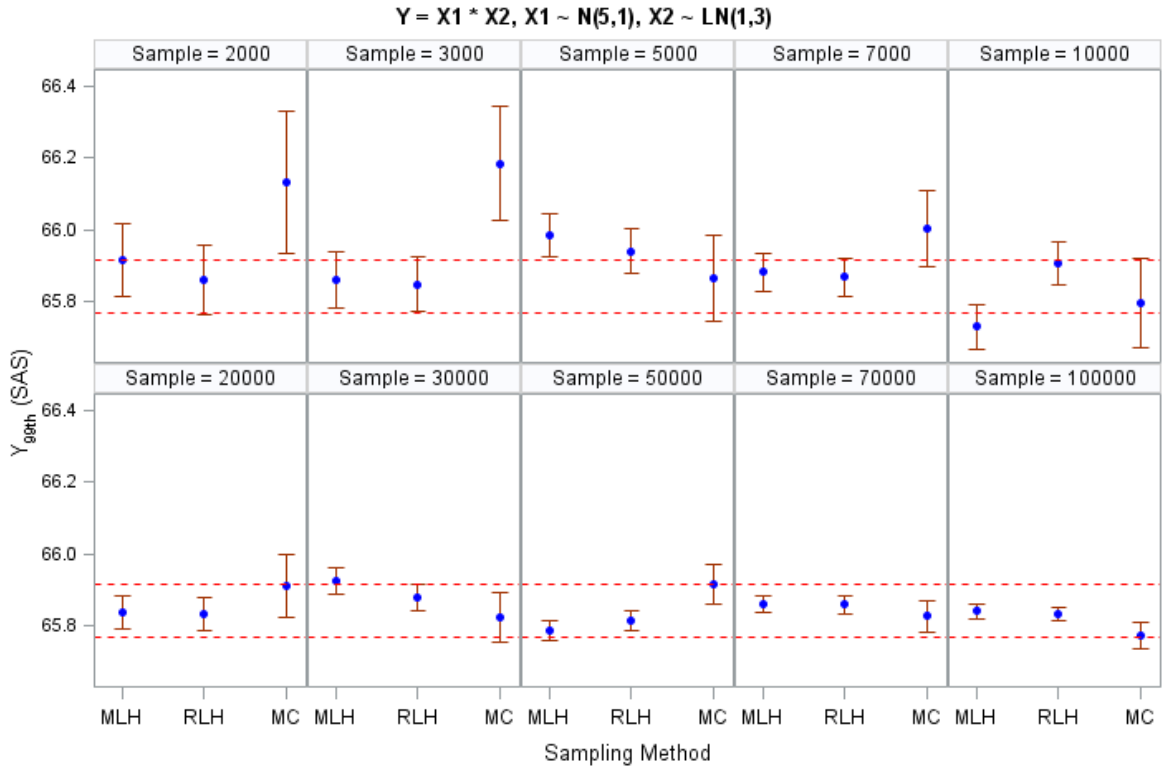


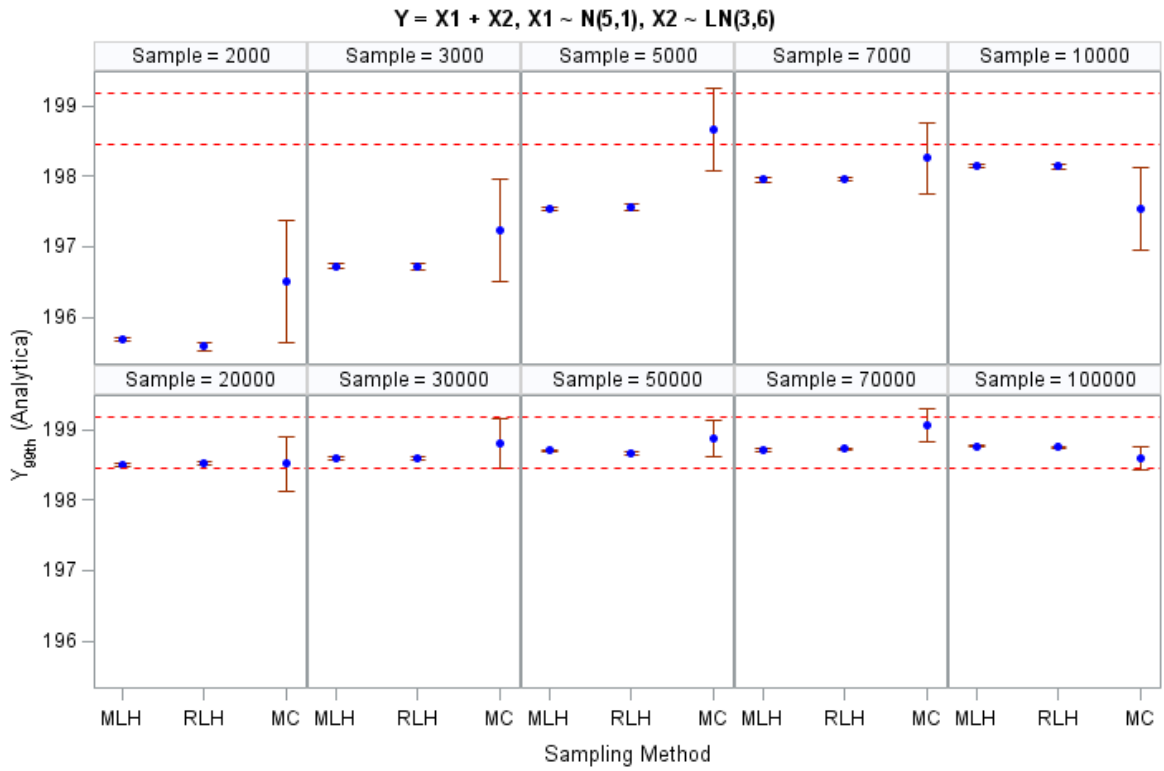


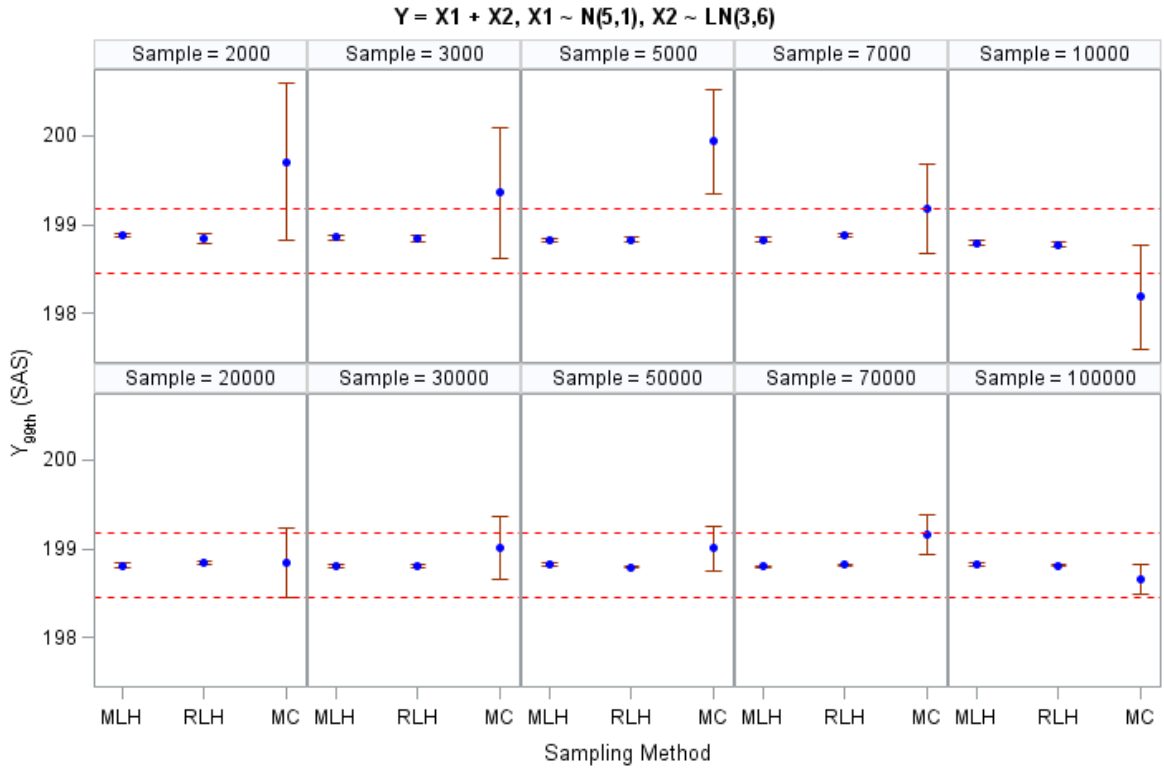


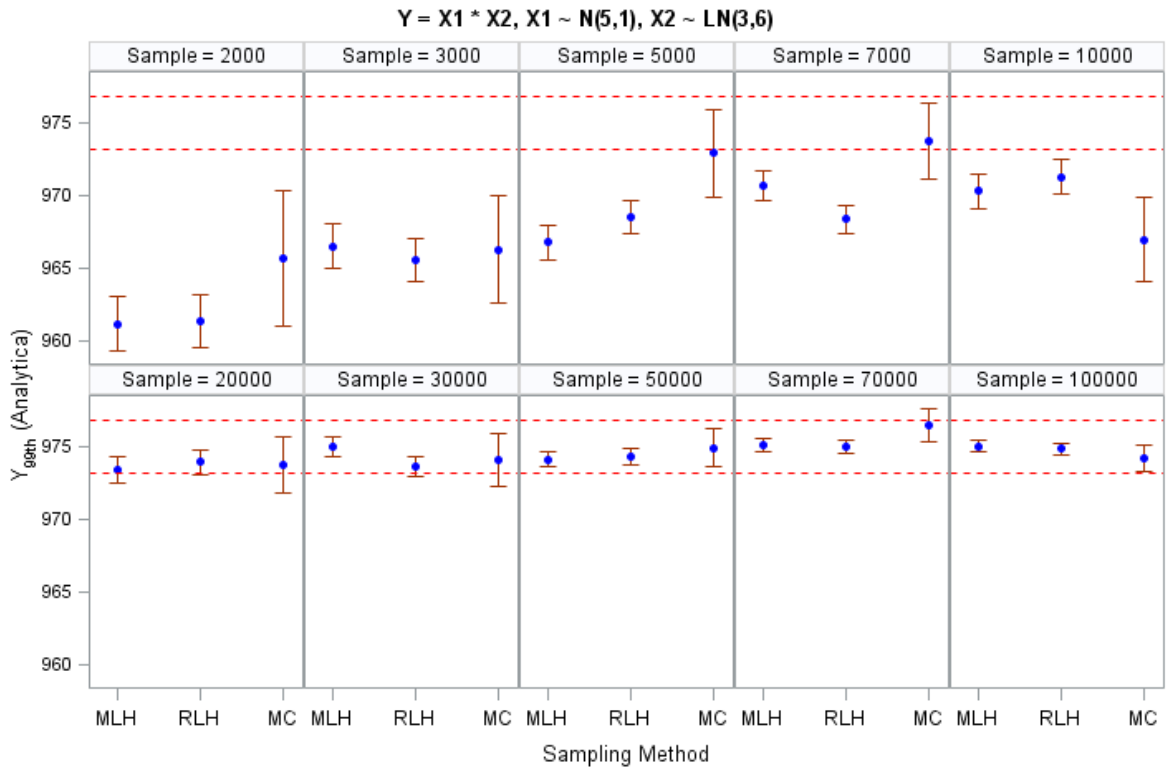


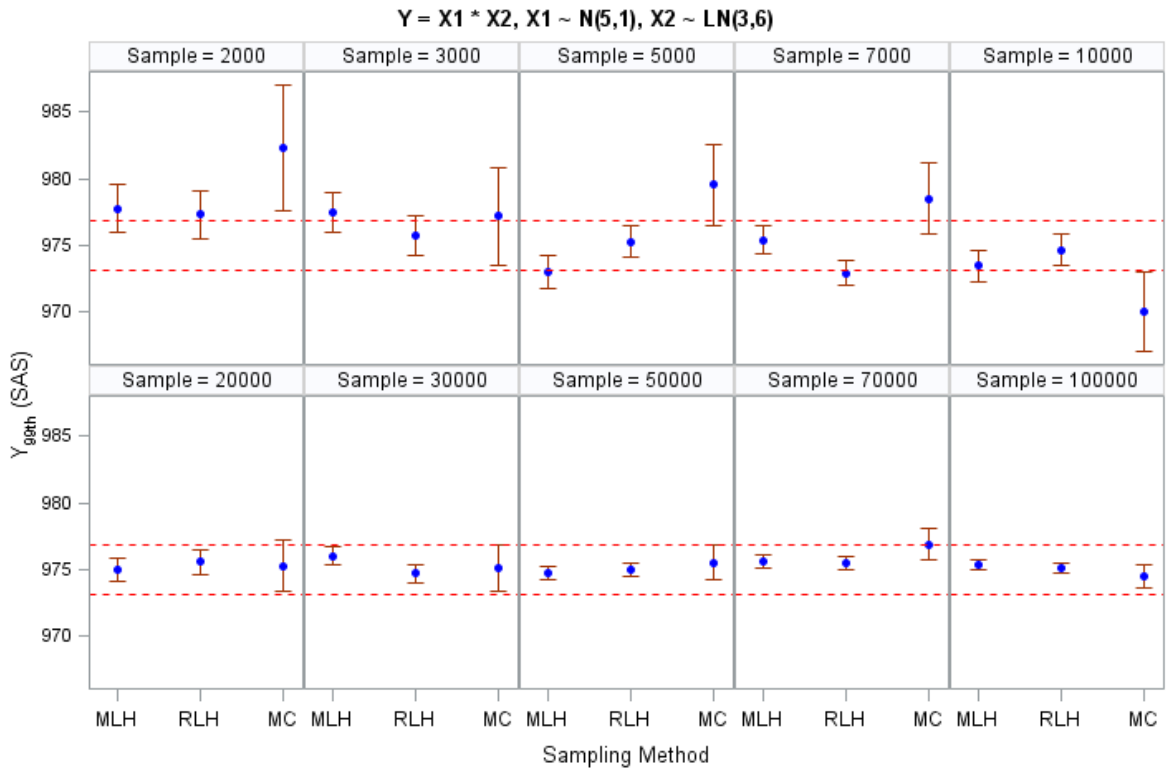


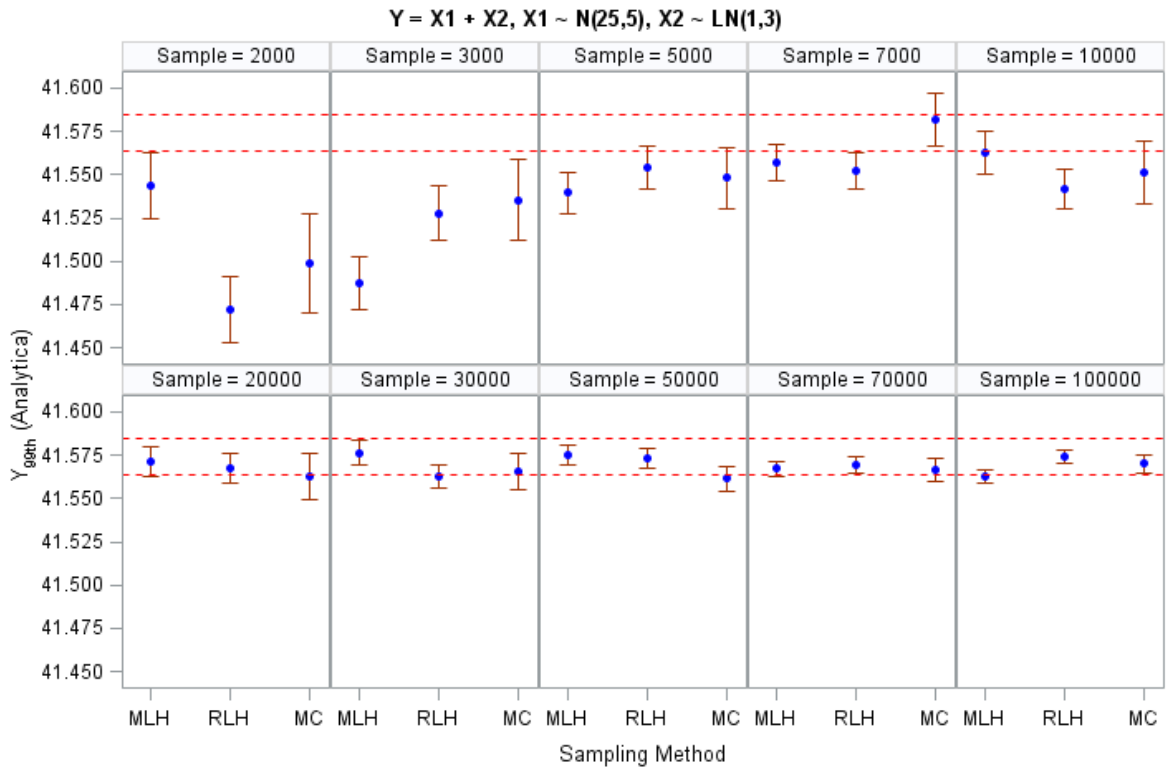


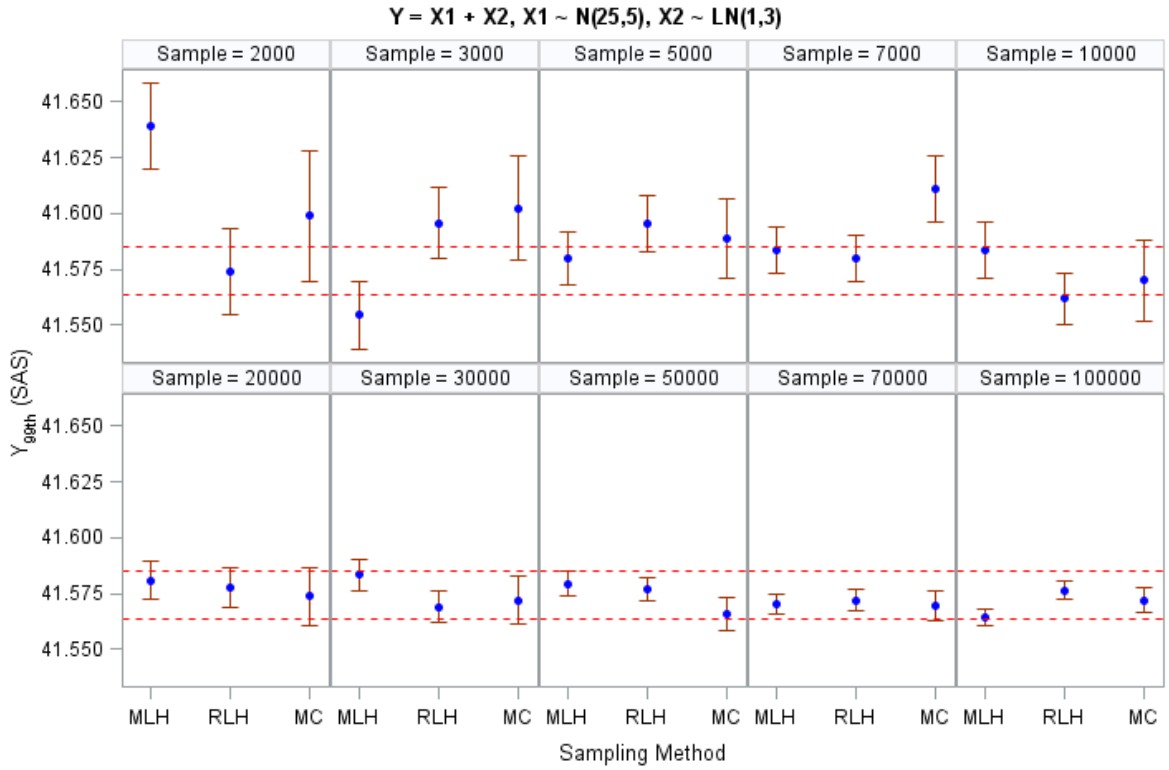


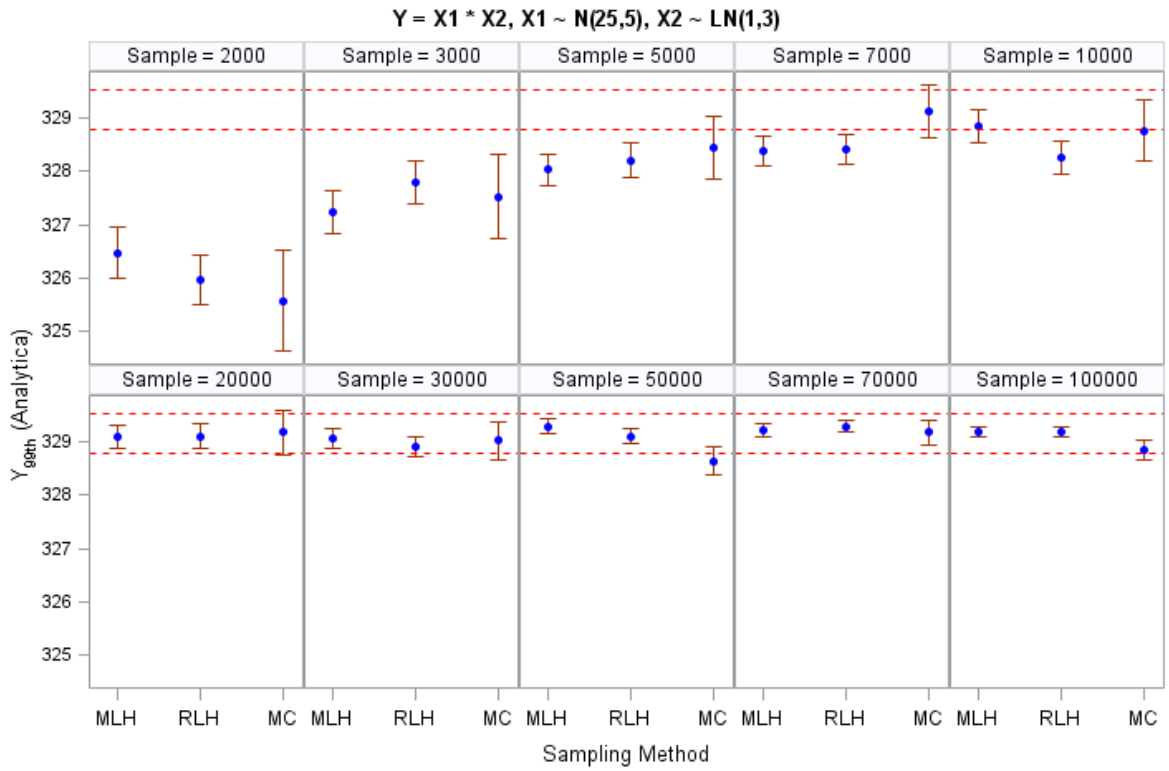


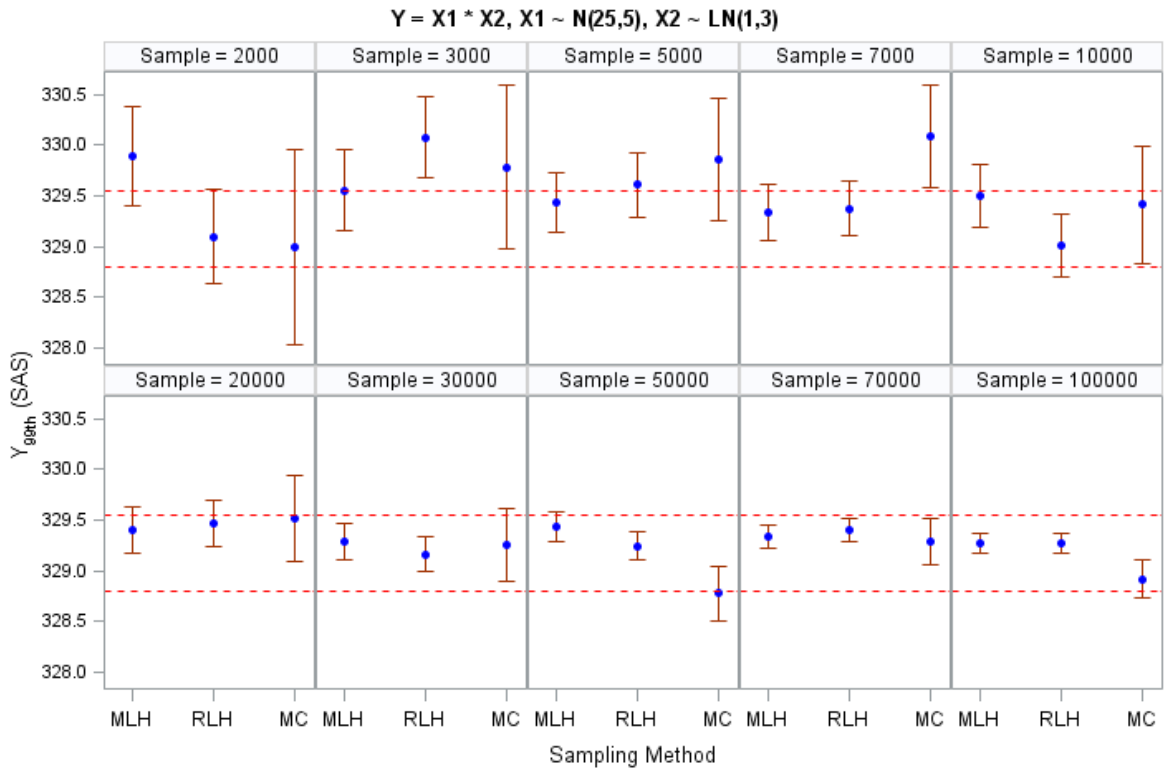


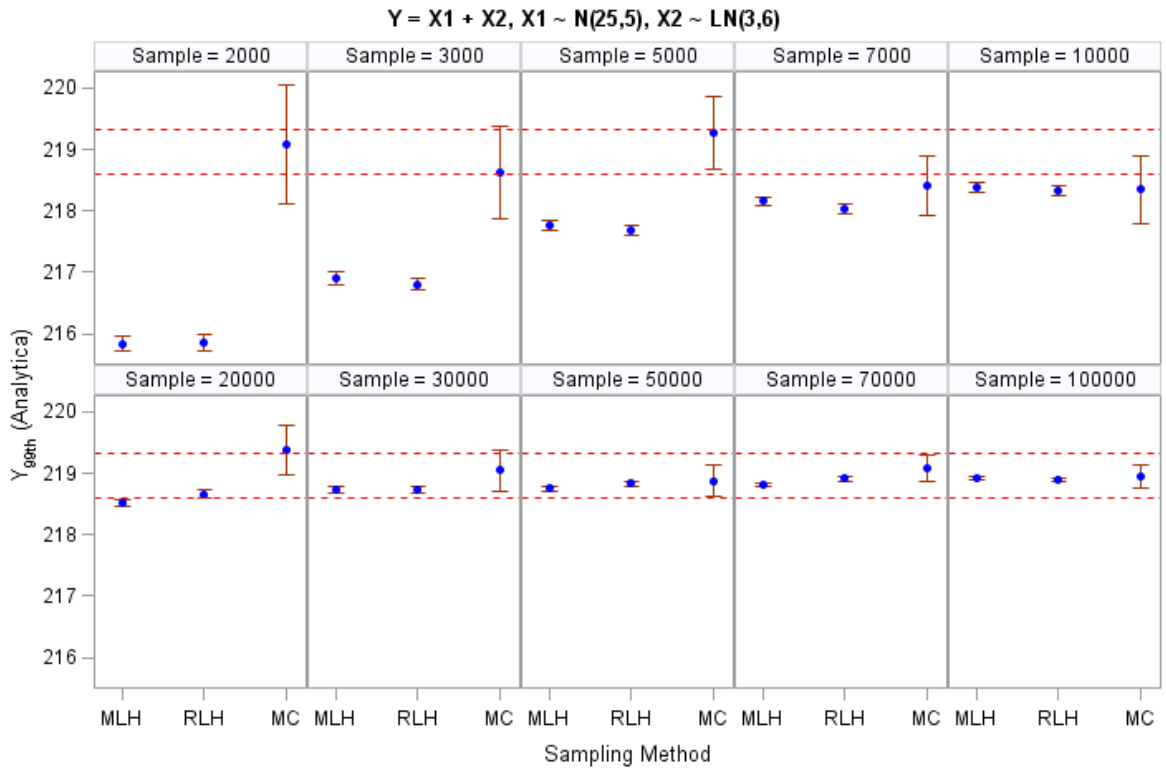


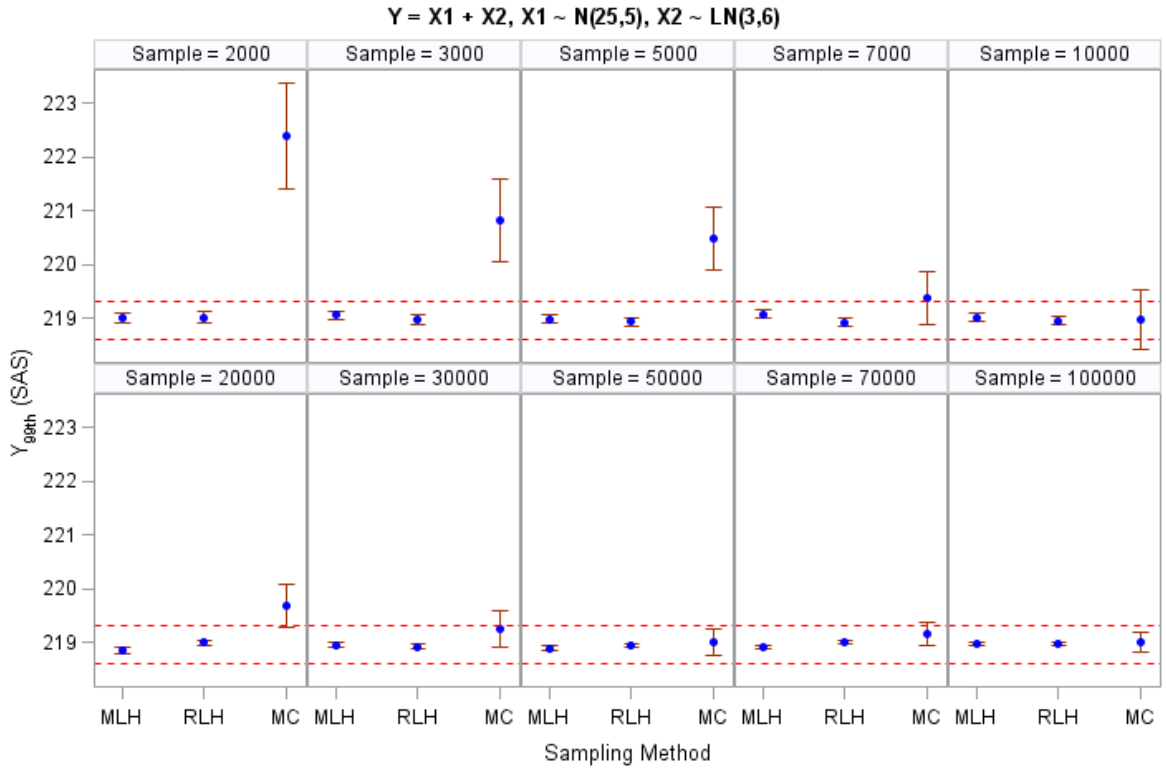


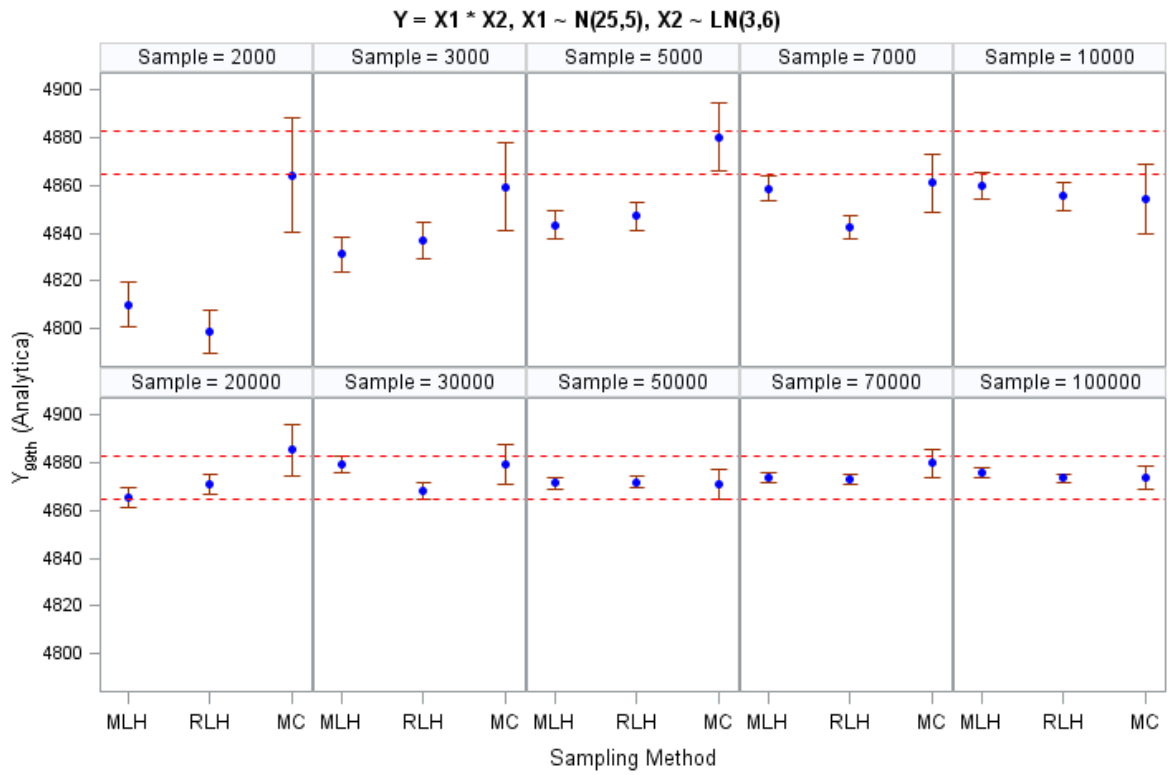


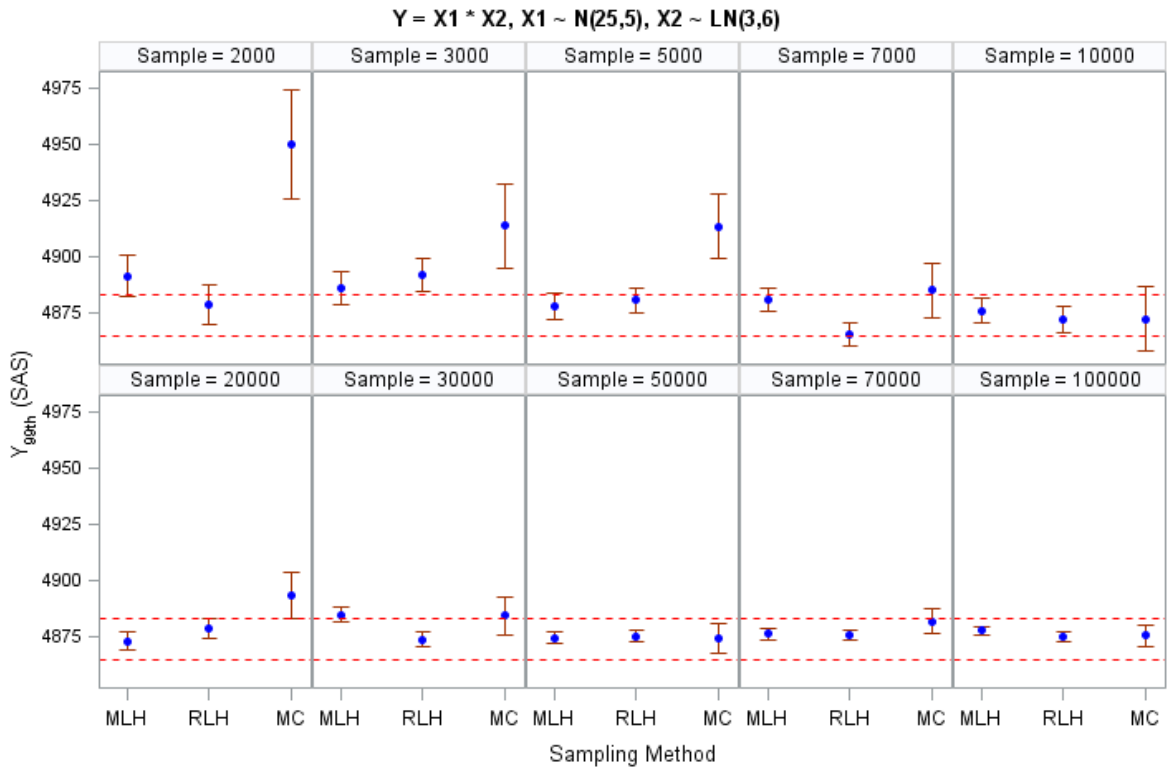












LIST OF TERMS, ABBREVIATIONS AND ACRONYMS

Sampling Methods

MLH	Median Latin Hypercube sampling method
RLH	Random Latin Hypercube sampling method
MC	Monte-Carlo sampling method

Distribution types

LN(GM, GSD)	Lognormal probability distribution defined by parameters GM and GSD
GM	Geometric Mean
GSD	Geometric Standard Deviation
N(mean, SD)	Normal probability distribution defined by parameters mean and SD
mean	Mean of a Normal (Gaussian) probability distribution
SD	Standard deviation
T(min,mode,max)	Triangular distribution defined by parameters min, mode and max
min	Minimum of a Triangular probability distribution
mode	Mode of a Triangular probability distribution
max	Maximum of a Triangular probability distribution
U(min,max)	Uniform distribution defined by parameters min and max
min	Minimum of a Uniform probability distribution
max	Maximum of a Uniform probability distribution
W(shape,scale)	Weibull distribution defined by parameters shape and scale
shape	shape parameter of a Weibull probability distribution
scale	scale parameter of a Weibull probability distribution

Other terms

99^{th} PC value	99^{th} percentile of probability of causation (PC)
ABRWH	Advisory Board on Radiation and Worker Health
IREP	Interactive Radio-Epidemiological Program
DCAS	Division of Compensation Analysis and Support
NIOSH	National Institute of Occupational Safety and Health
PC	Probability of Causation
RB	Relative Bias
RRMSE	Relative Root-Mean Square Error