Steady State Quantile Estimation

Mirko Eickhoff

Computer Science and Software Engineering,
University of Canterbury,
Christchurch, New Zealand,
m.eickhoff@cosc.canterbury.ac.nz

Abstract. Steady state simulation is used to study long-run behavior. Usually only the expected value of the steady state probability distribution function is estimated. In many cases quantiles of this distribution are of higher interest. In this paper a new usage of quantile estimators is proposed, which is derived from mean value analysis and is based on multiple independent replications. The advantage in using multiple independent replications is discussed, especially their ability to detect the steady state phase of quantiles.

1 Introduction

The purpose of steady-state simulation is to study the long-run behavior of a system. Using estimators for means, the results of the simulation can answer questions about the average system state like: How many customers are there on average in the queue? On the other hand, quantiles are known to be more robust against outliers than mean values. Quantile estimation can also answer questions like: What is the probability of more than k customers in the queue? Questions of this kind are often of more interest to the decision-maker. The complexity of quantile estimation is higher than that of mean value estimation, but the estimation of quantiles can give a deeper insight into the system of interest. This is true especially when several quantiles are estimated. A set of several quantiles can be used to estimate the steady-state distribution function. The estimation of the steady-state distribution is the ultimate goal in steady-state simulation.

The most important property of a quantile estimator is its statistical accuracy. The variance of a quantile estimator leads to errors, which usually decrease with an increase in the number of observations used for estimation. These errors are often referred to as random errors. They are caused by the fact that a stochastic measure is analysed and that every simulation is a statistical experiment. The next source of error is the bias of the estimator, which is often called the systematic error. This kind of error appears if e.g. assumptions about the analysed data hold only approximately. If both the variance and the bias tend to zero for a large number of observations the estimator is called consistent. More details about these statistical properties of a quantile estimator can be found in [18]. There are further properties besides these statistical ones, which characterize a suitable estimator. The storage requirements and the execution time are quite important, because usually a large number of output observations need to be processed to obtain trustworthy results. Therefore, not only the mathematical definition of the estimator, but also the method of computation is of interest. Efficient data structures and algorithms are important. To guarantee a proper use of the estimator in many situations, even for inexperienced users, it is important that the quantile estimator is easy to understand and that the number of user specified parameters is small, preferably zero. A classification of these properties is given in [13] for the general problem of estimating standard errors.

1.1 Single Quantile

The estimation of one quantile of a steady state distribution, when simulating a single instance (or "single replication") of a time-stationary process, is considered by Iglehart, Seila, Heidelberger and Lewis, Jain and Chlamtac, Chen and Kelton (see e.g. [17], [30], [16], [18] or the more recent article [4]). The methods of Igelhart and Seila are limited to regenerative processes. The subdivision of the output data into its regenerative cycles is a natural way to overcome the problem of autocorrelation. The method of Seila extends the method of Igelhart by grouping the regenerative cycles into batches. The number of parameters which have to be specified by the user is reduced by this batching approach to one parameter: the batch size. However, the determination of the batch size is a difficulty common to every batching approach; it is difficult for an inexperienced user to choose an appropriate value. The method of Heidelberger and Welch addresses the problem of quantile estimation in dependent sequences. Their method is not limited to regenerative processes. The point estimate based on ordered data is still valid in the dependent case, but its variance is inflated leading to a larger interval estimate. Two basic solutions are given. On the one hand, the higher variance can be calculated directly with the spectral method (see [15]). On the other hand, the data can be transformed to almost independent data by using a batch means method (see e.g. [11]). The method of Jain and Chlamtac uses a completely different kind of quantile estimator. Their estimator is based on markers, which are adjusted when collecting new observations. This is done by a piecewise-parabolic interpolation. Because of this interpolation, this method is not recommended for quantile estimation of discontinuous distribution functions. The estimator seems to be quite complicated compared to the usual estimators based on ordered data. However, the principal advantage is that the method requires only a constant (and small) amount of memory. Chen and Kelton describe a method that estimates a quantile by focusing on observations which are located in the neighbourhood of this quantile. Their method is sequential to ensure an accurate final estimate. However, the quality of this method has not been exhaustively studied yet.

A method for quantile estimation in finite-horizon simulation is described in e.g. [2]. This method is based on multiple replications of the finite-horizon simulation. These replications are dependent on each other because negative correlation is introduced into their streams of input random numbers to reduced variance. Avramidis and Wilson propose that this approach yields improvements under special assumptions (see also [19]).

The estimation of one single quantile is usually done to analyse the tail behaviour of a distribution. In this case typically the 0.95-quantile (resp. 0.05-quantile) is estimated. For more extreme quantiles than this it might be more appropriate to use rare event simulation. However, sometimes the median (0.5-quantile) is estimated instead of the mean value, because the median is more robust against outliers.

1.2 Several Quantiles

If the analyst is interested in the complete distribution function of a performance measure the estimation of several quantiles is useful, because the quantiles describe the probability distribution at special points. The estimation of several quantiles of the steady state distribution is addressed by Raatikainen (see [27]). The method of Jain and Chlamtac is extended by introducing additional markers to estimate more quantiles. The adjustment of the markers is done in the same way as before. An investigation of the variance of this method is given in [28].

One of the main difficulties in quantile estimation is the high computational effort and the large amount of storage needed to order the observations. Therefore, Heidelberger and Lewis reduce the sample size by a maximum transformation (see [16]). Jain, Chlamtac and Raatikainen go further and avoid sorting the output data by using an interpolation. In recent publications of Hashem, Schmeiser and Wood (see [14] and [35]) or Chen and Kelton (see [5] and [6]) quantile estimators based on order statistics have become popular again. This may be due to increased memory and processor speeds making these methods more practical. Wood and Schmeiser describe a batching method for quantiles which is similar to batch means and consider different quantile estimators, all based on ordered observations. The batch statistic is given by one of four quantile estimators, which are all based on ordered observations. Again, the difficulty is how to chose an appropriate batch size. In [5] the previous method of estimating a single quantile is extended to the problem of estimating several quantiles. Again, the extended method is sequential as the previous version.

1.3 Replications for Quantile Estimation

Very little is known about steady state quantile estimators based on multiple replications. We believe that the independence of the replications enables the use of new estimators with lower complexity and maybe higher performance. The use of multiple independent replications is discussed in the next section. In the following sections a new class of estimators is introduced, which are based on independent replications. In the last section we give conclusions of our research work so far.

2 Independent Replications

The main problem in quantile estimation for steady-state performance measures is that the output process $X_1, X_2, \ldots$ is typically not stationary and is autocorrelated, see e.g. [23]. Therefore, the number of required output data can be immense, which causes a problem when storing and sorting the output data. Using p independent replications of the simulation is a well known approach to obtain independent sequences of output data. Let $\left\{\{x_{j,i}\}_{i=1}^{n_j}\right\}_{j=1}^{p}$ denote the collected observations. $x_{j,i}$ is the ith observation of the jth replication. n_j denotes how many observations are collected in the jth replication. Collecting the same number of observations of each replications ensures that $\forall j : n_j = n$. Additionally, let assume that the ith observations of all replications describe the same measure. For example the ith observation could be the waiting time of the ith customer leaving a system. These assumptions ensure that the observations of the ith column are independent and identically distributed (iid).

$$\text{independent: } \Pr\left[\forall j : X_{j,i} \leq x\right] = \prod_j \Pr\left[X_{j,i} \leq x\right] \tag{1}$$

$$\text{identical: } \forall j : F_{X_{j,i}}(x) = F_{X_i}(x) \tag{2}$$

$F_{X_i}(x) = \Pr\{X_i \leq x\}$ denotes the cumulative probability distribution function (CDF) of a random variable X_i. $X_{j,i}$ is the random variable representing the ith observation in the jth replication. These properties help to overcome the main problem of quantile estimation in a single simulation run and enables the use of traditional quantile estimators for iid random samples of X_i.

$F_{X_i}(x)$ can be estimated by

$$\hat{F}_{X_i}(x) = \frac{1}{p}\sum_{j=1}^{p} \zeta(x - x_{j,i}) \tag{3}$$

with

$$\zeta(\Delta) = \begin{cases} 1 & \text{if} \quad \Delta \geq 0 \\ 0 & \text{otherwise} \end{cases} . \tag{4}$$

$\hat{F}_{X_i}$ is called the empirical cumulative distribution function (ECDF). If several values of $\hat{F}_{X_i}$ are of interest it is advisable to base the estimation on a sorted random sample. Let $\{y_{j,i}\}_{j=1}^{p}$ be the sorted values of $\{x_{j,i}\}_{j=1}^{p}$, so that $y_{j,i}$ is the jth order statistic. Based on these order statistics (3) can be transformed to

$$\hat{F}_{X_i}(x) = \frac{1}{p}\min(j | x \geq y_{j,i}) \tag{5}$$

with $1 \leq j \leq p$ and $\hat{F}_{X_i}(x) = 0$ for $x < y_{1,i}$. Note, that the value $y_{j,i}$ is an estimate of the j/p-quantile of F_{X_i}, compare Section 4.

Multiple independent replications enable the estimation of F_{X_i}. In [8] this is used to depict the transient behavior of the output process by plotting a suitable number of quantiles over time. Another application, which is based on the estimation of F_{X_i}, is described in [3] or [9] and is discussed in the next section.

3 Truncation Point Detection

To start a simulation experiment, the simulation model has to be initialized at an initial state I. This initial state has an impact on the random variable X_i and influences its CDF. Therefore, the initial state I has to be included in the CDF of X_i: $F_{X_i}(x|I) := \Pr[X_i \leq x|I]$. Assuming an ergodic system, $F_{X_i}(x|I)$ converges towards $F_X(x) = \lim_{j\to\infty} F_{X_i}(x|I)$ which is called the marginal CDF of the output process $\{X_i\}_{i=1}^{\infty}$ in steady state. The primary concern of steady state simulation is to determine this distribution or its specific measures, such as moments or quantiles.

In general the influence of I is significant at the beginning and decreases with increasing model time. If the interest is focused on the steady state behavior of the system, this initialization bias is obviously undesirable. A common way to reduce the influence of I is to truncate the "most" influenced part of the stochastic output process $X_1, \ldots, X_{l-1}$. Following this strategy the problem is to find an appropriate truncation point l. In the literature the steady state phase $\{X_j\}_{j=l}^{\infty}$ is described as a phase which is "relatively free of the influence of initial conditions" [12] or by the statement that $X_l, X_{l+1}, \ldots$ "will have approximately the same distribution" [22]. In practise there will often be an observation index l, such that

$$\forall i \geq l : F_{X_i}(x|I) \approx F_X(x) \tag{6}$$

holds, unless the process $\{X_i\}_{i=1}^{\infty}$ is statistically unstable. Of course l should be finite, and should be the minimum of all indices, for which (6) holds. Even though the estimation of $F_X(x)$ is the ultimate goal of steady state simulation, the expected value of the steady state random variable $E[X] = \lim_{j\to\infty} E[X_i]$ is often the only measure of interest. In this situation it is a generally accepted approach to replace (6) by

$$\forall i \geq l : E[X_i] \approx E[X] \ . \tag{7}$$

In general, however, the convergence of the expected value is only a necessary condition for stationarity, and not sufficient, see [34]. Therefore, (6) can be applied in analysis of arbitrary measures, especially for steady state quantile estimation. If ever, (7) should be used in mean value analysis only.

Finding a truncation point on the basis of (6) is not straightforward. It is therefore not very surprising that the most common methods for detection of the truncation point are based on (7), see [25], or on a visual inspection of transformed output data, see e.g. [34]. As already shown in the previous section, F_{X_i} can be estimated from multiple independent replications. A completely automated approach to detect a suitable value l is described in [9]. This heuristic assumes that at least $p = 30$ parallel replications are available. Its worst case time complexity is $\mathrm{O}(np\log(p))$, which is quite efficient, regarding that the number of collected observations is np.

3.1 Nonparametric Homogeneity Test

A very important point of this truncation point detection method is the homogeneity test used. In the goodness-of-fit problem the null hypothesis

$$H_0 : F_{X_0}(x) = F_{X_1}(x) = \cdots = F_{X_{k-1}}(x) \tag{8}$$

is checked by a homogeneity test. A k-sample version compares k random samples with each other. For our purpose, a nonparametric test with no further assumptions about F_{X_i} is needed.

The Kolmogorov-Smirnov test, see [20] and [32], is a nonparametric homogeneity test. In the 2-sample version the test statistic is

$$\mathsf{KS_2} = \sup_{-\infty < x < \infty} |\hat{F}_{X_0}(x) - \hat{F}_{X_1}(x)| \ , \tag{9}$$

where $\hat{F}_{X_0}(x)$ and $\hat{F}_{X_1}(x)$ are the ECDF of X_0 and X_1 consisting of p_0 resp. p_1 random values $x_{0,0}$, …, $x_{p_0,0}$ and $x_{0,1}$, …, $x_{p_1,1}$ in sorted order.

An algorithmic approach to calculate $\mathsf{KS_2}$ can be based on two pointers, which are shifted within the range of the random variable. By shifting these pointers in parallel through the interval $[\min(x_{0,0}, x_{0,1}), \max(x_{p_0,0}, x_{p_1,1})]$ the difference $\hat{F}_{X_0}(x) - \hat{F}_{X_1}(x)$ can be calculated for every value of x. Critical values for $\mathsf{KS_2}$ are known for different α-levels.

The Anderson-Darling test, see [1], is a nonparametric homogeneity test, like the Kolmogorov-Smirnov test. In [29] a k-sample version is introduced, which uses the test statistic

$$\mathsf{AD_k} = \sum_{i=0}^{k-1} p_i \int_{-\infty}^{\infty} \frac{(\hat{F}_{X_i}(x) - H'(x))^2}{H'(x)(1 - H'(x))} \mathrm{d}H'(x) \ . \tag{10}$$

p_i is the sample size of X_i and $H'(x)$ denotes the ECDF of the pooled sample of all X_i with $0 \le i \le k-1$. A computational formula for $\mathsf{AD_k}$ is given by

$$\mathsf{AD_k} = \frac{1}{N} \sum_{i=0}^{k-1} \frac{1}{p_i} \sum_{j=1}^{N-1} \frac{(NM_{ij} - jp_i)^2}{j(N-j)} \ , \tag{11}$$

where M_{ij} is the number of observations in the sample of X_i, which are smaller or equal than Z_j. $Z_1 < Z_2 < \cdots < Z_N$ denotes the pooled and ordered sample of $H'(x)$ with $N = \sum_{i=0}^{k-1} p_i$.

In [29] it is shown that $\mathrm{E}\left[\mathsf{AD_k}\right] = k - 1$ holds if all $F_{X_i}(x)$ are continuous and if the null hypothesis, see (8), can be assumed. To check the null hypothesis, additionally the variance of $\mathsf{AD_k}$ is needed. It is given by

$$\mathrm{Var}\left[\mathsf{AD_k}\right] = \frac{aN^3 + bN^2 + cN + d}{(N-1)(N-2)(N-3)} \ . \tag{12}$$

For details on the calculation of a, b, c and d see [29]. $\mathsf{AD_k}$ can now be standardized by

$$T_k = \frac{\mathsf{AD_k} - \mathrm{E}\left[\mathsf{AD_k}\right]}{\mathrm{Var}\left[\mathsf{AD_k}\right]} \ . \tag{13}$$

Critical values of T_k are tabulated for $k < 12$ and various α-level. If $k \geq 12$ holds, the critical value of T_k is given by

$$t_k = b_0 + \frac{b_1}{\sqrt{k-1}} + \frac{b_2}{k-1} \quad . \tag{14}$$

Again, values of b_0, b_1 and b_2 are tabulated for various α-level.

3.2 Accuracy

The most interesting performance measure of the truncation point detection method is its ability to estimate l. Our experience with previous implementations of this method ([3] and [9]) is that its accuracy is lower if the initial state I influences mostly the tail of the density function of $F_{X_i}(x|I)$. For example if the mean is constant but the variance is changing over time. We believe that this problem is introduced through the KS_2 statistic, which is based on the maximum difference.

To test whether the KS_2 or the AD_k (with $k = 2$) statistic delivers better results, we applied them on two artificial output processes with a well defined truncation point l:

$$X_t^{(A)} = \begin{cases} \epsilon_t + x - t\frac{x}{l} & \text{if } t < l, \\ \epsilon_t & \text{else.} \end{cases} \tag{15}$$

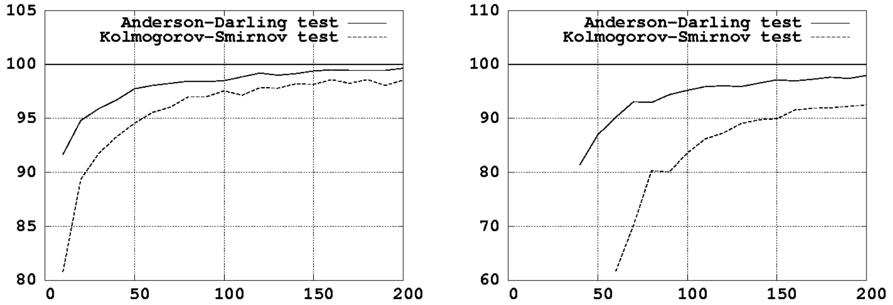
$$X_t^{(B)} = \begin{cases} \epsilon_t \cdot (x - t\frac{x-1}{l}) & \text{if } t < l, \\ \epsilon_t & \text{else.} \end{cases} \tag{16}$$

with $x = 10$, $l = 100$. The randomness is introduced by the Gaussian white noise process ϵ_t with the distribution $\mathrm{N}(0, 1)$. $X_t^{(A)}$ is governed by a transient mean value, whereas $X_t^{(B)}$ is governed by a transient variance. The results of the truncation point detection method are depicted in Figure 1. Experiments are done for various values of $p \leq 200$. The abscissa shows p, the number of parallel replications, and the ordinate shows the estimated truncation point l. It is clearly evident that for both processes the estimation of l based on AD_k is closer to the theoretical value $l = 100$.

3.3 Time Complexity

The results of the previous section clearly suggest the use of AD_k instead of KS_2. However, in practise the time complexity to calculate these statistics is another important performance measure. In the best case the calculation of AD_k should not require a higher computational effort. The worst case time complexity of both statistics is investigated next.

Theorem 1. *The worst case time complexity of the Kolmogorov-Smirnov 2-sample test is* $\mathrm{O}(N \log N)$ *with* $N = p_0 + p_1$.



(a) Transient Mean: $X_t^{(A)}$

(b) Transient Variance: $X_t^{(B)}$

Fig. 1. Performance of $\mathsf{KS_2}$ and $\mathsf{AD_k}$.

Proof. The random samples of X_0 and X_1 have to be sorted. Sorting of these two samples can be done in $\mathrm{O}(p_0 \log p_0 + p_1 \log p_1)$. Assuming that $p_0 > 0$ and $p_1 > 0$ the unequation $p_0 \log p_0 + p_1 \log p_1 < N \log N$ holds. Therefore, the execution time of sorting can be bounded by $\mathrm{O}(N \log N)$.

The calculation of the difference $|\hat{F}_{X_0}(x) - \hat{F}_{X_1}(x)|$ at a given value of x can be done in $\mathrm{O}(1)$, because only a constant number of basic arithmetic operations are involved. The algorithm is passing through the range of x by jumping from a $x_{i,j}$ to its successor in sorted order. Because there are $p_0 + p_1 = N$ values of x in total, the maximum difference can be calculated in $\mathrm{O}(N)$.

If the given samples are small, the critical value can be looked up in the given table in $\mathrm{O}(1)$. If the given samples are large, the critical value can be calculated by a constant number of basic arithmetic operations, which leads again to $\mathrm{O}(1)$. The comparison of the maximum difference and the critical value needs another $\mathrm{O}(1)$.

The summary of all results leads to $\mathrm{O}(N \log N) + \mathrm{O}(N) + \mathrm{O}(1) + \mathrm{O}(1)$. This shows, that the cardinal operation of the Kolmogorov-Smirnov 2-sample test is the sorting of the data. Therefore, the worst case execution time is $\mathrm{O}(N \log N)$. □

Theorem 2. *The worst case time complexity of the execution of a Anderson-Darling k-sample test is* $\mathrm{O}(N^2 + N \log N + kN)$ *with* $N = \sum_{i=0}^{k-1} p_i$.

Proof. Sorting the random samples of X_i can be done in $\mathrm{O}(p_i \log p_i)$. Consequently, sorting of all k random samples can be done in $\mathrm{O}(\sum_{i=0}^{k-1} p_i \log p_i)$. Because $\forall i (0 \leq i < k) : p_i > 0$ is valid, the overall sorting time can can be bounded by $\mathrm{O}(\sum_{i=0}^{k-1} p_i \log p_i) < \mathrm{O}(N \log N)$. By passing in parallel through all k sorted random samples the sequence $Z_1 < Z_2 < \cdots < Z_N$ can be generated. Each value has to be accessed only once, therefore, this can be done in $\mathrm{O}(\sum_{i=0}^{k-1} p_i) = \mathrm{O}(N)$. The ith column $\{M_{ij}\}_{j=1}^{N}$ of the M_{ij}-matrix can be calculated by passing in parallel through $\{Z_j\}_{i=j}^{N}$ and the ith sorted random sample. This is done in $\mathrm{O}(N + p_i)$. Processing all k columns leads to a run time of $\mathrm{O}(kN + \sum_{i=0}^{k-1} p_i) = \mathrm{O}(kN + N) = \mathrm{O}(kN)$. If the M_{ij}-matrix is known, the calculation of the fraction in (11) is done in $\mathrm{O}(1)$, because a constant number of arithmetic operations are needed. The inner sum of that equation loops over $N-1$ values and the outer sum loops over k values. Therefore, the calculation of (11) needs $k(N-1) \cdot \mathrm{O}(1) = \mathrm{O}(kN)$ steps. Combining all previous results leads to $\mathrm{O}(N \log N) + \mathrm{O}(N) + \mathrm{O}(kN) + \mathrm{O}(kN) = \mathrm{O}(N \log N + kN)$, which is the overall worst case execution time to calculate the test statistic $\mathsf{AD_k}$.

To normalize the test statistic $\mathsf{AD_k}$ its variance is needed. Here, the calculation of the parameters a, b, c and d, see (12), is not discussed in detail. However, the cardinal equation to calculate these parameters is

$$\sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \frac{1}{(N-i)j} \ , \tag{17}$$

see [29]. The calculation of the fraction in (17) is done in $\mathrm{O}(1)$. Both, the inner sum and the outer sum of that equation loop over maximum $N-2$ values. Therefore, (17) can be calculated in $\mathrm{O}((N-2)^2) \cdot \mathrm{O}(1) = \mathrm{O}(N^2)$ steps. The calculation of T_k, see (13), can now be done with a constant number of basic arithmetic operations in $\mathrm{O}(1)$. Therefore, the complete normalization can be done in $\mathrm{O}(N^2)$.

The critical value t_k can be calculated in $\mathrm{O}(1)$, no matter of the value of k. Because in every case a constant number of tabled values and basic arithmetic operations are needed. Combining all results, the overall run time of the Anderson-Darling k-sample test is given by $\mathrm{O}(N \log N + kN) + \mathrm{O}(N^2) + \mathrm{O}(1) = \mathrm{O}(N^2 + N \log N + kN)$. □

The test statistic $\mathsf{AD_k}$ depends on the difference of the ECDFs. In contrast to the $\mathsf{KS_2}$ statistic not only the maximum difference is used, but the integral resp. sum over the whole range of x. The higher computational complexity is caused by the calculation of $\mathrm{Var}\,[\mathsf{AD_k}]$, which is not depending on the data itself, but on the size of the random samples. If many Anderson-Darling tests on random samples of constant size are performed $\mathrm{Var}\,[\mathsf{AD_k}]$ has to be calculated only once. This is exactly the situation when performing the truncation point detection algorithm on the output data of multiple replications, because $\forall j : p_j = p$. The dominating factor in the calculation of the $\mathsf{AD_k}$ statistic itself is the sorting of the data. Therefore, the time complexity of the truncation point detection method remains the same, no matter whether the $\mathsf{KS_2}$ or the $\mathsf{AD_k}$ statistic is used. The use of $\mathsf{AD_k}$ involves an additional calculation time of $\mathrm{Var}\,[\mathsf{AD_k}]$ before the simulation is started. Compared to the whole run time of this method, the additional calculation time is negligible.

The empirical investigation shows that the truncation point detection based on the statistic $\mathsf{AD_k}$ is more accurate. Thus, the $\mathsf{KS_2}$ statistic, which is used in [8], should be replaced by $\mathsf{AD_k}$.

4 Quantile Estimators for Independent Replications

The q-quantile of F_{X_i} is defined by $q = F_{X_i}(x(q,i))$ and, therefore,

$$x(q,i) = F_{X_i}^{-1}(q) = \inf\{x | F_{X_i}(x) \geq q\} \tag{18}$$

is the location of the q-quantile in the case of a continuous distribution $F_{X_i}(x)$ with $0 \leq q \leq 1$. As before, let $\{y_{j,i}\}_{j=1}^{p}$ be the ordered values of $\{x_{j,i}\}_{j=1}^{p}$. A valid estimator for the location of the q-quantile at observation index i is given by

$$\hat{x}(q,i) = y_{\lceil pq \rceil, i} \ . \tag{19}$$

To avoid rounding of non integer values, it is suitable to choose $q = j/p$ with $j = \{1; 2; \cdots ; p\}$.

In the previous section an approach to detect a valid truncation point l is described, so that (6) is fulfilled. The homogeneity test ensures that the difference in distribution among the remaining X_i is negligible. In consequence, all $\{y_{j,i}\}_{i=l}^{\infty}$ describe the j/p-quantile of the steady state probability distribution F_X.

$$\hat{x}(q) = \frac{1}{n-l+1} \sum_{i=l}^{n} \hat{x}(q,i) \tag{20}$$

is a point estimate of $F_X^{-1}(q)$. By performing p independent replications we obtain p quantiles of F_X at $q = j/p$ with $j = \{1; 2; \cdots ; p\}$. They are equally distributed within the probability domain $[0; 1]$.

Theorem 3. *$\hat{x}(q)$ is an unbiased estimator of $F_X^{-1}(q)$ for large p and l.*

Proof. The expected value of (20) is

$$\begin{aligned} \mathrm{E}\left[\hat{x}(q)\right] &= \frac{1}{n-l+1} \sum_{i=l}^{n} \mathrm{E}\left[\hat{x}(q,i)\right] \qquad (21) \\ &= \frac{1}{n-l+1} \sum_{i=l}^{n} \mathrm{E}\left[y_{\lceil pq \rceil, i}\right] \ . \end{aligned}$$

$\mathrm{E}\left[y_{\lceil pq \rceil, i}\right] = F_{X_i}^{-1}(q)$ holds for large values of p (see [7] or [6]). Furthermore, all $X_l, X_{l+1}, \ldots$ are assumed to be identically distributed, i.e. $\forall i : F_{X_i}(x) = F_X(x)$. Equation (21) evaluates to

$$\begin{aligned} \mathrm{E}\left[\hat{x}(q)\right] &= \frac{1}{n-l+1} \sum_{i=l}^{n} F_{X_i}^{-1}(q) \qquad (22) \\ &= \frac{1}{n-l+1} \sum_{i=l}^{n} F_X^{-1}(q) \\ &= F_X^{-1}(q) \ . \end{aligned}$$

The estimator $\hat{x}(q)$ is asymptotically unbiased, i.e. $\mathrm{E}\left[\hat{x}(q)\right] - F_X^{-1}(q) = 0$, because (22) holds for large p and l. □

Every simulation is a statistical experiment. Point estimators never return exact values, even if they are unbiased. Confidence intervals (or interval estimates) are essential to provide convincing results. To establish a confidence interval for (20) its variance $\mathrm{Var}\left[\hat{x}(q)\right]$ is needed. Note, that $\{y_{j,i}\}_{i=l}^{\infty}$ (row) is autocorrelated and the variance cannot be estimated directly. The form of estimator (20) is identical to mean value estimators of single simulation runs. Here the specialty is that each component describes the j/p-quantile. Therefore, known techniques for variance estimation of mean value estimators can be applied. Spectral analysis and batching methods are commonly used in mean value analysis.

4.1 Spectral Analysis

In [15] a confidence interval for the steady state mean value is generated by spectral analysis on basis of a single simulation run. This confidence interval is used to control run length to obtain estimates with a specified accuracy. This method assumes, that the output sequence converges to a steady state behaviour which can be modeled as a covariance stationary process. The problem of the initial transient phase is not addressed, but the correlation of the output data. This approach is originally used for mean value analysis. In conjunction with the maximum transformation, it is also used for quantile estimation of one single quantile, see [16].

The sequence $\{y_{j,l}, y_{j,l+1}, \ldots, y_{j,n}\}$ fulfils the precondition for this spectral method, even though the analysed measure is the j/p-quantile and not the mean. Furthermore, the sequence starts with the lth observation so that the problem of the initial transient can be neglected at this stage. The spectral method of Heidelberger and Welch seems to be applicable in this context. We describe now how spectral analysis can be used to establish a confidence interval for the point estimator (20).

Let $\{y_{j,i}\}_{i=l}^{n}$ be realization of the stochastic process $\{Y_{j,i}\}_{i=l}^{n}$. The covariance function $\gamma(k)$ is defined by

$$\gamma(k) = \mathrm{Cov}\left[Y_{j,i}, Y_{j,i+k}\right] \quad . \tag{23}$$

Because the process is assumed to be covariance stationary the absolute value of i does not matter, as long as $l \leq i$. The spectral density $\rho(f)$ at frequency f can be derived by

$$\rho(f) = \sum_{k=-\infty}^{+\infty} \gamma(k) \cos(2\pi f k) \quad . \tag{24}$$

According to Heidelberger and Welch,

$$\mathrm{Var}\left[\hat{x}(j/p)\right] = \frac{\rho(0)}{n - l + 1} \tag{25}$$

is valid if $n - l + 1$ is large. This means that a confidence interval for (20) can be constructed if $\rho(f)$ can be estimated at $f = 0$.

The periodogram $I(m/N)$ with $N = n - l + 1$ is defined by

$$I(m/N) = \frac{|\sum_{j=l}^{n} Y_{j,i} e^{-2\pi\sqrt{-1}(i-1)m/N}|^2}{N} \quad . \tag{26}$$

It helps to estimate $\rho(m/N)$ because

- $\mathrm{E}\left[I(m/N)\right] = \rho(m/N)$,
- $\mathrm{Var}\left[I(m/N)\right] = \rho^2(m/N)$ and
- $\mathrm{Cov}\left[I(m/N), I(m'/N)\right] = 0$

hold for $0 < m < N/2$ and $m \neq m'$. If some points of $I(m/N)$ are known, a polynomial fit can be used to estimate $\rho(0)$. However, in [15] is demonstrated that the use of $I(m/N)$ is not optimal for this purpose because the variance of $I(m/N)$ is not constant and its skewness is not zero. The transformation to

$$J(f_m) = \log\left(\frac{I\left((2m-1)/N\right) + I\left(2m/N\right)}{2}\right) \tag{27}$$

with $f_m = (4m-1)/(2N)$ leads to more convenient properties:

- $\mathrm{E}\left[J(f_m)\right] = \log\left(\rho(f_m)\right) - 0.27$,
- $\mathrm{Var}\left[J(f_m)\right] = 0.645$ and
- $\mathrm{Cov}\left[J(f_m), I(f_{m'})\right] = 0$

hold for $0 < m < N/2$ and $m \neq m'$. Furthermore, the skewness of $J(f_m)$ is about zero.

The polynomial fit is based on two parameters. The first parameter K is the number of points of $J(f_m)$ used to obtain the polynomial fit. The second parameter d is the degree of the polynomial. In [15] an algorithm is given to estimate $\hat{\rho}(0)$. In [24] the standard setting $d = 2$ and $K = 25$ of this algorithm is discussed. A positive slope of $\hat{\rho}(f)$ at $f = 0$ can lead to a too small estimate of $\rho(0)$. By using the maximum of $\hat{\rho}(0)$ for $d = 0$ or $d = 2$ the coverage of the confidence interval is increased. Finally, a confidence interval can be derived by assuming that

$$\frac{\hat{x}(q) - F_X^{-1}(q)}{\sqrt{\frac{\hat{\rho}(0)}{n-l+1}}} \tag{28}$$

is governed by a t-distribution.

Using order values of p independent replications, as described in Section 2, p output processes $\{y_{j,i}\}_{i=l}^{\infty}$ are available and (20) can be applied for all $q = j/p$ with $1 \leq j \leq p$. $\hat{x}(j/p)$ and $\mathrm{Var}\left[\hat{x}(j/p)\right]$ can be calculated for all j separately, as well as the confidence intervals based on $\mathrm{Var}\left[\hat{x}(j/p)\right]$. By calculating several quantiles we receive an estimate of the steady state probability distribution F_X.

4.2 Batching Method

The literature about batching methods is vast. Possibly one of the earliest described batching methods in conjunction with simulation output analysis is [10]. The basic idea is to divide the output process into subsequences, called batches, of equal size. For all batches a batch statistic is calculated, e.g. the batch mean. The use of this approach is that the batch statistics become approximately independent for a large batch size. The assumed independence helps to estimate the variance of the batch statistics. The difficulty of this method is the determination of an appropriate batch size.

To keep it simple, we use non-overlapping and non-disjoint batches. The transformed data is given by

$$z_{j,i}(m) = \frac{1}{m} \sum_{k=1}^{m} y_{j,(l-1+im-k)} \tag{29}$$

with $1 \leq j \leq p$, $1 \leq i \leq n'$ and $n' = \frac{n-l+1}{m}$. The size of the resulting data matrix is reduced by $1/m$. The interval estimator $\hat{x}(q)$ can now be calculated by

$$\hat{x}(q) = \frac{1}{n-l+1} \sum_{i=l}^{n} y_{\lceil pq \rceil,i} = \frac{1}{n'} \sum_{i=1}^{n'} z_{\lceil pq \rceil,i}(m) \ . \tag{30}$$

This equation is a different denotation of (20), because the sum over the batch means is calculated. With an appropriate choice of m the batch means $z_{\lceil pq \rceil,1}(m)$, $z_{\lceil pq \rceil,2}(m)$, ... are approximately independent of each other. Under this assumption $\mathrm{Var}\left[\hat{x}(q)\right]$ can be estimated, by

$$\sigma^2_{\hat{x}(q)} = \frac{1}{n'(n'-1)} \sum_{i=1}^{n'} \left(z_{\lceil pq \rceil,i}(m) - \hat{x}(q) \right)^2 \tag{31}$$

as it is done in [10]. $(\hat{x}(q) - F_X^{-1}(q))/\sigma_{\hat{x}(q)}$ is approximately t-distributed with n' degrees of freedom, thus a confidence interval can be constructed. For every $q = j/p$ with $1 \leq j \leq p$ the expected value $\mathrm{E}\left[\hat{x}(q)\right]$, its variance $\mathrm{Var}\left[\hat{x}(q)\right]$ and the belonging confidence interval can be estimated. We receive p interval estimates of quantiles of the steady state distribution $F_X(x)$.

To estimate confidence intervals for all $q = j/p$ an overall batching approach can be performed, which operates on $\{y_{j,i}\}_{i=l}^{\infty}$ for all j in parallel. Again, the determination of an appropriate overall batch size m, which is valid for all j, is the bottleneck. Tests for independence based on runs, see [31], or lag-1 autocorrelation, see [21], are used to detect a valid batch size. Most lag-1 autocorrelation test assume the normality of the distribution, which is only approximately true. For small sample sizes complex corrections of the test statistic are done, see e.g. [11]. Therefore, we introduce a heuristic test, which is based on milder assumptions and promises good performance in our context of determining the overall batch size m.

Median Confidence Interval for Pearson's Correlation Coefficient. Let $\hat{r}^{(p)}(P_1)$ be Pearson's correlation coefficient of the original lag-1 paired batch means $\{(z_{j,i}(m); z_{j,i+1}(m))\}_{i=1}^{n'-1}$. And let $\hat{r}^{(p)}(P_k)$ be Pearson's correlation coefficient for the lag-1 paired data of the kth permutation of $\{z_{j,i}(m)\}_{i=1}^{n'}$ with $2 \leq k \leq (n'!)$. In [26] the first four moments of Pearson's correlation coefficient are derived. Here, the first and the third moment are of special interest: $\mathrm{E}\left[\hat{r}^{(p)}\right] = 0$ holds even for small samples and $\mathrm{Skew}\left[\hat{r}^{(p)}\right] = 0$ holds approximately. $\mathrm{Skew}\left[\hat{r}^{(p)}\right]$ defines the degree of asymmetry of the distribution of $\hat{r}^{(p)}$. Therefore, $F_{\hat{r}^{(p)}}(0) = 0.5$ is approximately true. The null hypothesis of our test is that $\{z_{j,i}(m)\}_{i=1}^{n'}$ is an independent sequence.

$$\Pr\left[|\hat{r}^{(p)}(P_k)| < |\hat{r}^{(p)}(P_1)|\right] = \frac{1}{2} \tag{32}$$

holds under the null hypothesis and for a randomly chosen permutation P_k. For K randomly chosen permutations $P_{k_1}, \ldots, P_{k_K}$ we can derive

$$\Pr\left[\forall l (1 \leq l \leq K) : |\hat{r}^{(p)}(P_{k_l})| < |\hat{r}^{(p)}(P_1)|\right] = \frac{1}{2^K} \ . \tag{33}$$

On base of this equation a confidence interval can be established:

$$\Pr\left[-\Delta \leq \hat{r}^{(p)}(P_1) \leq \Delta\right] = 1 - \frac{1}{2^K} \tag{34}$$

with halfwidth

$$\Delta = \max_{1 \leq l \leq K} \left(|\hat{r}^{(p)}(P_{k_l})|\right) \ . \tag{35}$$

If $\hat{r}^{(p)}(P_1)$ is not within the confidence interval, the null hypothesis must be rejected at confidence level $1 - \frac{1}{2^K}$. For a general discussion on median confidence intervals see [33]. The advantage of using this confidence interval is that the assumption of zero skewness is milder than the assumption of a normal distribution. For only $K = 6$ permutations the confidence level is already > 0.95 and K can be regarded as a constant parameter. Therefore, the time complexity of the confidence interval calculation is the same as for the calculation of $\hat{r}^{(p)}(P_1)$ itself. For our purpose of estimating the overall batch size m for p independent replications this correlation test is performed on $\{z_{j,i}(m)\}_{i=1}^{\infty}$ for any j.

5 Future Work

An unsolved issue is the selection of a stopping rule for sequential estimation of quantiles. Stopping rules for mean value analysis are based on the relative error. The situation for quantiles is slightly different, especially if several quantiles are accessed in parallel. Our ultimate goal is to estimate the steady state probability distribution based on quantiles.

After solving this issue performance measures of the described quantile estimators can be investigated and compared. The most important measure is the statistical accuracy, which can be determined e.g. by coverage analysis of the interval estimators. Empirical investigations of test models with known steady state distributions are necessary. Another important measure is the algorithmic complexity, involving the analysis of the needed computation time and storage requirements.

6 Conclusion

A survey of the recent state of the research work on steady state quantile estimation is given in the introduction. The advantage of the use of multiple independent replications is discussed. The suitability of multiple replications for steady state quantile estimation is pointed out.

In contrast to other methods, the use of multiple replications enables the detection of the steady state phase based on the empirical cumulative distribution function. The statistical accuracy of the truncation point detection method is increased by using a homogeneity test based on the Anderson-Darling test statistic. The computational complexity remains the same, because only miner calculations are added in the beginning of the simulation experiment.

A new point estimator for quantiles is proposed, which is based on ordered values of multiple replications. In addition two ways of constructing confidence intervals for this point estimator are discussed, which are derived from mean value analysis.

One of the introduced interval estimators involves the determination of an overall batch size for the ordered output streams describing different quantiles. For this purpose a test of independence is proposed by calculating median confidence intervals for Pearson's correlation coefficient. This test is based on mild assumptions.

7 Acknowledgement

I would like to thank K. Pawlikowski and D. McNickle for giving me direction and advice whenever I need it. They contributed to this paper by sharing their experience and time with me.

References

1. T. W. Anderson, D. A. Darling: A Test of Goodness of Fit. Journal of the American Statistical Association, Vol. 49, pp. 765-769, 1954
2. A. N. Avramidis and J. R. Wilson: Correlation-Induction Techniques for Estimating Quantiles in Simulation Experiments. Operations Research, Vol. 46, No. 4, pp. 574-591, July-August 1998
3. F. Bause and M. Eickhoff: Truncation Point Estimation using Multiple Replications in Parallel. Proceedings of the Winter Simulation Conference, pp. 414-421, 2003
4. E. J. Chen and W. D. Kelton: Simulation-Based Estimation of Quantiles. Proceedings of the Winter Simulation Conference, pp. 428-434, 1999

5. E. J. Chen and W. D. Kelton: Quantile and Histogram Estimation. Proceedings of the Winter Simulation Conference, pp. 451-459, 2001
6. E. J. Chen: Two-Phase Quantile Estimation. Proceedings of the Winter Simulation Conference, pp. 447-455, 2002
7. H. A. David: Order Statistics. John Wiley & Sons, Inc., 1970
8. M. Eickhoff, D. McNickle and K. Pawlikowski: Depiction of Transient Performance Measures using Quantile Estimation. Proceedings of the 19th European Conference on Modelling and Simulation (ECMS'2005), pp. 358-363, 2005
9. M. Eickhoff, D. McNickle and K. Pawlikowski: Efficient Truncation Point Estimation for Arbitrary Performance Measures. Proceedings of the 3rd Industrial Simulation Conference (ISC'2005), pp. 5-12, 2005
10. G. S. Fishman: Grouping Observations in Digital Simulation. Management Science, Vol. 24, No. 5, pp. 510-521, January 1978
11. G. S. Fishman and L. S. Yarberry: An Implementation of the Batch Means Method. INFORMS Journal on Computing, Vol. 9, No. 3, pp. 296-310, Summer 1997
12. G. S. Fishman: Discrete-Event Simulation. Springer, 2001
13. D. Goldsman and B. W. Schmeiser: Computational Efficiency of Batching Methods. Proceedings of the Winter Simulation Conference, pp. 202-207, 1997
14. S. Hashem and B. W. Schmeiser: Algorithm 727 Quantile Estimation Using Overlapping Batch Statistics. ACM Transactions on Mathematical Software, Vol. 20, No. 1, pp. 100-102, March 1994
15. P. Heidelberger and P. D. Welch: A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations. Communications of the ACM, Vol. 24, No. 4, pp. 233-245, April 1981
16. P. Heidelberger and P. A. W. Lewis: Quantile Estimation in Dependent Sequences. Operations Research, Vol. 32, No. 1, pp. 185-209, February 1984
17. D. L. Igelhart: Simulating Stable Stochastic Systems, VI: Quantile Estimation. Journal of the Association for Computer Machinery, Vol. 23, No. 2, pp. 347-360, April 1976
18. R. Jain and I. Chlamtac: The P^2 Algorithm for Dynamic Calculations of Quantiles and Histograms without Storing Observations. Communications of the ACM, Vol. 28, No. 10, pp. 1076-1085, October 1985
19. X. Jin, M. C. Fu and X. Xiong: Probabilistic Error Bounds for Simulation Quantile Estimators. Management Science, Vol. 14, No. 2, pp. 230-246, February 2003
20. A. N. Kolmogorov: Confidence Limits for an Unknown Distribution Function. Annals of Mathematical Statistics, Vol. 12, No. 4, pp. 461-463, 1941
21. A. M. Law and J. S. Carson: A Sequential Procedure for Determining the Length of a Steady-State Simulation. Operations Research, Vol. 27, No. 5, pp. 1011-1025, September-October 1979
22. A. M. Law and W. D. Kelton: Simulation Modeling and Analysis. McGraw-Hill Higher Education, New York, 2000
23. J-S. R. Lee, D. McNickle and K. Pawlikowski: Quantile Estimation in Sequential Steady-State Simulation. Proceedings of the 13th European Simulation Multiconference, pp. 168-174, 1999
24. D. McNickle, G. Ewing and K. Pawlikowski: Refining Spectral Analysis for Confidence Interval Estimation in Sequential Simulation. Proceedings of the 16th European Simulation Symposium, 2004
25. K. Pawlikowski: Steady-state simulation of queueing processes: a survey of problems and solutions. ACM Computing Surveys, Vol. 22, pp. 123-170, June 1990

26. E. J. G. Pitman: Significance Tests Which May be Applied to Samples from any Populations. II. The Correlation Coefficient Test. Supplement to the Journal of the Royal Statistical Society, Vol. 4, No. 2, pp. 225-232, 1937
27. K. E. E. Raatikainen: Simultaneous estimation of several percentiles. SIMULATION, Vol. 49, No. 4, pp. 159-164, October 1987
28. K. E. E. Raatikainen: Sequential Procedure for Simultaneous Estimation of Several Percentiles. Transactions of the Society for Computer Simulation, Vol. 7, No. 1, pp. 21-44, 1990
29. F. W. Scholz, M. A. Stephens: K-Sample Anderson-Darling Tests. Journal of the American Statistical Association, Vol. 82, No. 399, pp. 918-924, September 1987
30. A. F. Seila: A Batching Approach to Quantile Estimation in Regenerative Simulations. Management Science, Vol. 28, No. 5, pp. 573-581, May 1982
31. S. Siegel: Nonparametric Statistics. McGRAW-HILL BOOK COMPANY, INC., 1956
32. N. V. Smirnov: Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics, Vol. 19, No. 2, pp. 279-281, June 1948
33. J. C. Strelen: The Accuracy of a new Confidence Interval Method. Proceedings of the 2004 Winter Simulation Conference, pp. 654-662, 2004
34. P. D. Welch: The Statistical Analysis of Simulation Results. In The Computer Performance Modeling Handbook, ed. S. Lavenberg, Academic Press, pp. 268-328, 1983
35. D. C. Wood and B. Schmeiser: Consistency of Overlapping Batch Variances. Proceedings of the Winter Simulation Conference, pp. 316-319, 1994