RUN LENGTH CONTROL FOR SIMULTANEOUS ESTIMATION OF SEVERAL PERCENTILES IN DEPENDENT SEQUENCES[†]

Kimmo E. E. Raatikainen
University of Helsinki, Department of Computer Science

Abstract: Percentiles are convenient indices to describe simulation output data. A few percentiles are more informative than the mean and the standard deviation. However, percentiles are very seldom used in practice. The reason has been that percentile estimation is regarded as a cumbersome task. All previously reported methods estimate only a single percentile. In addition, they have been methods of fixed, predefined, run length.

The method introduced in this paper does not have these limitations. It estimates several percentiles simultaneously without storing and sorting the output sequence. It also has a run length control mechanism, which terminates the simulation run, when the estimates are accurate enough. The relative half-width criterion, which is commonly used to control the accuracy of the simulated mean, is generalized to control the accuracy of multidimensional estimates.

The proposed method dynamically approximates the inverse of the empirical cumulative distribution function by a piecewise parabola. The run length control mechanism applies the coefficients of the parabola and the spectral method to estimate the variances and covariances of the percentiles. This technique assumes that the output sequence can reasonably be modeled as a covariance stationary strongly ϕ-mixing stochastic process. It also relies on the fact that the approximated percentiles are very near to the percentiles calculated from the order statistic. In an earlier paper the last assumption has been validated.

In the emprical studies, seven different processes are analyzed. Although the method applies several approximations, the results indicate that the proposed method can estimate the percentiles within the predefined accuracy. In fact, the accuracy is essentially better than required, which leads to quite long simulation runs.

[†]This work was supported by Emil Aaltonen Foundation.

Proc. Int. Conf. on Methodology and Validation (April 6–9, 1987; Orlando, Fla) SCS

1. INTRODUCTION

Percentiles are convenient measures to describe simulation output data. They have become very popular tools in the so called exploratory data analysis. The popularity is mainly due to robustness. A few exceptional observations do not disturb the percentiles as heavily as they disturb the mean and the standard deviation. A few percentiles provide a good summary of the entire range of output values. Especially a set of five percentiles — 5% or 10%, 25%, 50%, 75% and 95% or 90% — together with the extreme values has been found to be practical.

Percentiles have only seldom been used in simulation studies. We belive that the reason is the complexity of percentile estimation. The calculation of percentile estimates has been regarded as a cumbersome task. In discrete event simulation the accuracy estimation is not a simple task, since the output sequence is, in general, autocorrelated. Sophisticated approaches are necessary to obtain an accuracy estimate.

Two different approaches have been proposed to overcome the difficulties due to autocorrelation. The first approach is to assume that the output sequence is generated by a renewal process. Three regenerative methods falling in this class have been published. A good summary is given in [Seila 1982]. The second approach is based on the assumption that the output sequence has the ϕ-mixing property. This property quarantees that the percentiles have a (marginal) normal limiting distribution. Three different methods, which rely on the ϕ-mixing property, were introduced and analyzed in [Heidelberger and Lewis 1984].

Our method, which is introduced in this paper, tries to make the percentile estimation easier. The method has two properties, which are pioneering: the ability to estimate several percentiles simultaneously and the mechanism of controling the accuracy of the estimates. We apply data reduction heavily so that the amount of storage is small and fixed.

The method has three basic components. The first one estimates the percentiles and their marginal densities dynamically without storing and sorting the observations. The second component generates the associated binary processes and reduces them by batching. The third component is used to evaluate the accuracy measure of the estimates. It generates the covariance matrice of the estimates from the marginal densities and the cross-spectrum of the associated binary processes. The measure is proportional to the largest eigenvalue of the covariance matrice in relative metrics.

The observations are collected in blocks. The first two components of the method are used after each new block to update their information. The third component is only used in selected control points, where the accuracy of the estimates is measured. It is not practical to measure the accuracy after each new block. In fact we double the sample size between adjacent control points.

The estimation of the percentiles is done by a piecewise parabola, which approximates the inverse of the empirical cumulative distribution function. This technique, known as the P^2-algorithm, was introduced in [Jain and Chlamtac 1985]. In [Raatikainen 1986] the algorithm was extended to estimate several percentiles simultaneously. In this paper we add the covariance estimation and the run length control mechanism to the *extended* P^2-algorithm given in [Raatikainen 1986].

The estimation of the covariances is based on a similar approach as the spectral method in [Heidelberger and Lewis 1984]. However, we consider the problem of estimating several percentiles simultaneously. Our method requires that the output sequence can be reasonably modeled as a covariance stationary, strongly ϕ-mixing stochastic process. The stronger version of the ϕ-mixing property is necessary for the asymptotical multinormality of the estimates. Although our percentile estimates are not based on the order statistics, we apply their asymptotical properties to our estimates. The comparison of the order statistics and the *extended* P^2-algorithm given in [Raatikainen 1986] suggests that serious errors are not introduced.

The target level of the accuracy of the estimates is given as a pair: *probability* (α) and *maximum allowed relative distance from the true values* (ε). The interpreption is that the relative error of the estimate is greater than ε only with probability α. Our measure of accuracy takes into account the variances and correlations of the estimates. The decision criterion to terminate the simulation or to continue it generalizes the relative half-length of the confidence interval, which is commonly used to control the accuracy of one-dimensional estimates.

The accuracy requirement we adopted makes it necessary to estimate also the correlations between the estimates. A simpler solution would have been to omit the effects of the correlations. However, we are sure that the effects of the correlations must be taken into account. Otherwise the results would not be accurate enough, since the percentile estimates are positively correlated even in independent samples. In discrete event simulation the output sequence usually has a very high level positive autocorrelation, which dies slowly. This phenomenon increases the correlations between the estimates considerably.

In empirical studies, we analyzed seven stochastic processes, whose marginal distributions and autocorrelation structures are known. At least in these cases the method has excellent accuracy. Perhaps the accuracy is 'too good', since the simulation runs are quite long. However, we are not able to compare the run lengths to other methods, since we have not found any other method, which estimates several percentiles simultaneously and tries to achieve a predefined accuracy. Although the run lengths are long, their variation is fortunately low. The empirical results — the achieved accuracy, the low variation of the run lengths and the correlation structures of the relative errors — indicate that our approach is of practical importance.

In Section 2 we derive the accuracy requirement for multidimensional estimates and introduce the spectral theory, which is used in calculation of the covariances of the estimates. The proposed method is described in Section 3. The ideas used in the *extended* P^2-algorithm and the estimation methodology of the covariances are introduced. A detailed description of the *extended* P^2-algorithm is given in the Appendix. In Section 4 we give results of emprical studies and make the conclusions.

2. THEORETICAL BACKGROUND

The theoretical background of the method will be introduced in this section. First the accuracy requirement of a multidimensional measurement will be derived. The requirement is a straightforward generalization of the one used for the one-dimensional measurements. The spectral theory, on which the applied covariance estimator is based, will be described in the second part. The description is only informative. The proofs can be found in the given reference.

2.1 Accuracy Requirement for Multidimensional Measurements

For one-dimensional measurements, such as the mean or a single percentile point, the accuracy requirement is usually given as the maximum relative half-length of the confidence interval. Using this requirement the simulation is continued until

$$\Pr\{|x - \theta|/|x| < \varepsilon\} > 1 - \alpha\,, \tag{1}$$

where x is the measurement, θ is its true (unknown) value, ε is the maximum relative half-length of the confidence interval, and $1 - \alpha$ is the confidence level.

The above criterion can be generalized for multidimensional measurements. The condifence interval is replaced by the confidence region. The relative half-length must be changed to relative distance from the midpoint of the confidence region. With these extensions the termination criterion has the form

$$\Pr\{d(\underline{x}, \underline{\theta}) < \varepsilon^2\} > 1 - \alpha\,, \tag{2}$$

where $\underline{x}$ is the m-dimensional measurement, $\underline{\theta}$ is its true (unknown) value, $d(\underline{x}, \underline{\theta})$ is the squared relative distance: $\sum_{i=1}^{m}(1 - \theta_i/x_i)^2$, ε is the maximum allowed relative distance from the midpoint of the confidence region, and $1 - \alpha$ is the confidence level.

Suppose that the m-dimensional measurement $\underline{x}$ has a normal distribution with expectation $\underline{\theta}$ and covariance matrice Σ. Let S be the estimate of Σ so that fS has the Wishart distribution $W_m(f, \Sigma)$. (The Wishart distribution is the multivariate generalization of the χ^2-distribution.) The boundaries of the confidence region is then defined by the hyperellipsoid

$$(\underline{x} - \underline{\theta})' S^{-1} (\underline{x} - \underline{\theta}) < \frac{mf}{f - m + 1} F_{m, f+1-m; 1-\alpha}\,, \tag{3}$$

where $F_{f_1, f_2; x}$ is the $100x^{\text{th}}$ percentile point of the F-distribution with f_1 and f_2 degrees of freedom.

The maximum squared relative distance, $d(\underline{x}, \underline{\theta})$, when $\underline{\theta}$ is on the boundary defined by equation (3), is given by

$$\lambda_{max} \frac{mf}{f-m+1} F_{m,f-m+1;1-\alpha} , \tag{4}$$

where λ_{max} is the largest eigenvalue of the matrice $X^{-1}SX^{-1}$, and $X = \mathrm{diag}(x_1, \ldots, x_m)$.

Hence, the termination criterion (2) is equal to the condition

$$\lambda_{max} \leq \frac{f-m+1}{f} \frac{\varepsilon^2/m}{F_{m,f+1-m;1-\alpha}} . \tag{5}$$

In one-dimensional case the condition (5) reduces to

$$\hat{s}^2/x^2 < (\varepsilon/t_{f,1-\alpha/2})^2, , \tag{6}$$

where $\hat{s}^2$ is the estimated variance of x (with f degree of freedom), and $t_{f,1-\alpha/2}$ is the $100(1-\alpha/2)^{\text{th}}$ percentile point of the Student's t- distribution with f degree of freedom.

The condition (6) is the same as that derived from the maximum relative half-width criterion.

2.2 Asymptotical Properties of the Order Statistics

It is shown in [Sen 1972] that percentile estimates, based on the order statistics, have a multinormal limiting distribution, if the following three conditions are satisfied:

1. The process $\{X_n\}$ satisfies the strong ϕ-mixing condition.
2. The (marginal) cumulative distribution function, $F(x)$, is absolutely continuous.
3. The (marginal) density function, $f(x)$, is finite, positive and absolutely continuous for all x for which $0 < F(x) < 1$.

Let x_p and x_q denote the $100p^{\text{th}}$ and $100q^{\text{th}}$ percentile, and $\hat{x}_p = X_{(\lfloor np \rfloor + 1)}$ and $\hat{x}_q = X_{(\lfloor nq \rfloor + 1)}$. Then

$$\lim_{n\to\infty} n\mathrm{Cov}(\hat{x}_p, \hat{x}_q) = \frac{h_{pq}(0)}{f(x_p)\, f(x_q)} , \tag{7}$$

where $h_{pq}(0)$ is the cross-spectrum of the binary processes $\{I_j(x_p)\}$ and $\{I_j(x_q)\}$ at frequency 0, and $I_j(x) = 1$, if $X_i \leq x$, otherwise $I_j(x) = 0$.

The cross-spectrum $h_{pq}(\omega)$ is, in general, complex valued. However, at frequency 0 the imaginary part is zero. Hence in estimation of $h_{pq}(0)$ one could omit the imaginary part. This is not a good approach, since the imaginary part affects the sampling properties of the real part. The approach we found to be the most practical one is based on the the following form of the norm of the cross-spectrum at frequency ω:

$$|h_{pq}(\omega)| = \sqrt{h_p(\omega)\, h_q(\omega)}\; |w_{pq}(\omega)| \,, \tag{8}$$

where $h_p(\omega)$ and $h_q(\omega)$ are the auto-spectrums, and $|w_{pq}(\omega)|$ is the coherency of the processes.

Thus the covariance matrice of the percentile estimates $\hat{x}_{p1}, \ldots, \hat{x}_{pm}$ can be approximated as

$$S = DWD \,, \tag{9}$$

where $D = \mathrm{diag}(\sqrt{\hat{h}_{p_1}(0)/n}\,/\hat{f}(\hat{x}_{p_1}), \ldots, \sqrt{\hat{h}_{p_m}(0)/n}\,/\hat{f}(\hat{x}_{p_m}))$, and $(W)_{kj} = e_{kj}\,|\hat{w}_{kj}(0)|$, where e_{kj} is either +1 or -1 depending on the sign of $\hat{h}_{p_k p_j}(0)$.

The comparison of the percentile estimates, based on the proposed method and the order statistics, is reported in [Raatikainen 1986]. The comparison shows that the estimates have very similar behavior. Hence, we can apply the results given above in the proposed method.

3. DESCRIPTION OF THE METHOD

In this section the proposed method will be described. The description will be given in two parts. In the first part, the method of estimating the percentiles and densities is introduced. This method is a straightforward generalization of the P^2-algorithm, which was published in [Jain and Chlamtac 1985]. The second part contains the estimation of the associated binary processes and their cross-spectrum.

3.1 Estimation of Percentiles and Densities

The method of estimating the percentiles is called the *extended* P^2-algorithm. It was introduced in [Raatikainen 1986]. The method is a straightforward generalization of the original P^2-algorithm given in [Jain and Chlamtac 1985]. The original algorithm was designed to estimate a single percentile. The extended version can estimate several percentiles simultaneously. The ideas of the extended algorithm are described in this section. The Appendix contains the algorithm in all details.

Instead of storing and sorting the observations, the P^2- algorithm updates the 'marker information' after each observation. Each marker has a height, a position and a desired position, and is associated with the estimation of a specific percentile.

Suppose that a marker is used to estimate the percentile at point p, and a sample of n observations $\{x_i\}_1^n$ is available. Let $x_{(j)}$ denote the j^{th} observation in the ordered sample. Now the desired position of the marker is $1 + (n-1)p$. The algorithm tries to keep the actual position of the marker, n_a, close to $[1+(n-1)p]$, which is the nearest integer to $1+(n-1)p$. The height approximates the value of $x_{(n_a)}$ and is thus an estimate of the percentile at point p.

Suppose that percentiles at points $p_1, \ldots, p_m$ are to be estimated. The algorithm consists of maintaining m+2 principal markers and m+1 middle markers. The principal markers are used to estimate the minimum, the required percentiles, and the maximum. Each middle marker estimates the percentile at the point midway between two adjacent principal markers.

The markers are updated, if necessary, after each observation. After n observations, the actual and desired positions of the markers at the minumum and at the maximum are 1 and n, respectively. Their heights are the minimum and the maximum values in the sample.

The updating of the actual and desired positions of the other markers is simple. For each marker, whose height is greater than the observation, the actual position is increased by one. In addition, if the actual position of a marker is now off to the left or to the right of its desired position by more than one position, then the height and the actual position are adjusted. The adjustment of the marker height is done using either a piecewise-parabolic or a piecewise-linear formula. The parabolic formula is preferable, but sometimes the linear formula must be used in order to keep the marker heights in a nondecreasing order.

The parabolic formula assumes that the curve passing through any three adjacent markers is of form $q_i = an_i^2 + bn_i + c$, where q_i is the height and n_i is the actual position of the marker. The movement of the actual position of a marker is always only one position. The parabolic adjustments are

$$\begin{aligned} q_i \leftarrow q_i \quad &+ \quad \frac{d}{n_{i+1}-n_{i-1}}\Big((n_i - n_{i-1} + d)\frac{q_{i+1}-q_i}{n_{i+1}-n_i} + \\ &\qquad (n_{i+1} - n_i - d)\frac{q_i - q_{i-1}}{n_i - n_{i-1}}\Big)\,, \\ n_i \leftarrow n_i \quad &+ \quad d\,, \end{aligned} \tag{10}$$

where $d = 1$, if the movement is to the right, and $d = -1$, if the movement is to the left.

The linear adjustments are

$$\begin{aligned} q_i &\leftarrow q_i + d\frac{q_{i+d}-q_i}{n_{i+d}-n_i}\,, \\ n_i &\leftarrow n_i + d\,, \end{aligned} \tag{11}$$

The piecewise-parabolic formula for the heights is actually an approximation of the inverse of the empirical cumulative distribution function, $\hat{F}^{-1}(y) = ay^2 + by + c$, in the neighbourhood of p. Hence the density function can be approximated by $\hat{f}(x) = (b + 2a\hat{F}(x))^{-1}$ in the neighbourhood of x_p.

The initialization of the algorithm is done by generating a block of n_b observations. These observations are sorted. The actual positions are initialized to $n_i = [d_i]$, where d_i is the desired position of the marker i, $i = 1, \ldots, 2m+3$. Then it is checked that the actual positions are strictly increasing. If not, some adjustments must be made. The heights are then initialized to $q_i = x_{(n_i)}$.

3.2 Estimation of the Associated Binary Processes

It was shown in Section 2.2 that the associated binary processes, $\{I_j(x_p)\}$, have a central role, when the covariances of the percentiles are estimated. In principle, the generation of these processes requires that the whole sequence $\{X_j\}$ is available. However, if the binary processes are accurately generated, the advantages of the extended P^2-algorithm are lost.

The problem of the generation of the associated binary processes is solved through approximating the processes $\{I_j(x_p)\}$ by the processes $\{\hat{I}_j(\hat{x}_p)\}$. The observations are generated in blocks containg n_b items. After each block we first update the percentile estimates using the extended P^2-algorithm. Let the current block be the k^{th} one and let $\hat{x}_{p,k}$ denote the estimated $100p^{\text{th}}$ percentile after kn_b observations. The items $\{\hat{I}_j(\hat{x}_p)\}_{j=1}^{(k-1)n_b}$ are not changed but n_b new items are generated as

$$\hat{I}_j(\hat{x}_p) = \begin{cases} 1\,, & \text{if } X_j \le \hat{x}_{p,k} \\ 0\,, & \text{otherwise} \end{cases} \quad , \; j = 1 + (k-1)n_b, \ldots, kn_b\,. \tag{12}$$

To avoid the storage of $\{\hat{I}_j(\hat{x}_p)\}_1^n$, the information is batched. The sequences $\{\hat{I}_j(\hat{x}_p)\}_1^n$ are divided dynamically into a fixed number (M) of subsequences, whose means, $\{\tilde{I}_j(\hat{x}_p)\}_1^M$, are stored. The batching is relevant, since it can be shown that $h_n(0)/n \approx h_M(0)/M$, where $h_n(\omega)$ is the cross-spectrum of the original sequences and $h_M(\omega)$ is the cross-spectrum of the batched sequences.

The covariance of $\hat{x}_{p_j}$ and $\hat{x}_{p_k}$ can be approximated as

$$\hat{s}^2_{p_j,p_k} = \frac{\sqrt{\hat{h}_{p_j}(0)/M}}{\hat{f}(\hat{x}_{p_j})} \frac{\sqrt{\hat{h}_{p_k}(0)/M}}{\hat{f}(\hat{x}_{p_k})} \, e_{p_j p_k} \left|\hat{w}_{p_j p_k}(0)\right| , \tag{13}$$

where the autospectrums, the coherency and the sign of the cross-spectrum are estimated from the batched sequences.

A convenient method to estimate the autospectrums at frequency 0 has been developed in [Heidelberger and Welch 1981]. A second order polynomial is fitted to the logarithms of the smoothed periodogram ordinates. The periodogram ordinates are evaluated from the Fourier coefficients of the sequences. In [Heidelberger and Lewis 1984] the method was applied to estimate the variance of percentile estimates.

In our method we apply directly the spectral method given in [Heidelberger and Welch 1981] to estimate the autospectrums. The estimation of the coherency is done using a similar approach.

First the cross-periodogram ordinates are evaluated from the Fourier coefficients. The ordinates are then smoothed by averaging every three adjacent values. The smoothed ordinates are then used to evaluate the coherency estimates. Let $|\hat{w}_k|$ denote the k^{th} estimate, $k = 1, \ldots, M/6 - 1$. The next step is the Fisher's z-transformation, $z_k = \tanh^{-1}(|\hat{w}_k|)$, which stabilize the variance of $|\hat{w}_k|$ and render it independent of frequency. Then a second-order polynomial is fitted to the points $\{(3k-1)/M, z_k - 1/4\}$. The estimate $|\hat{w}(0)|$ is finally obtained as $\tanh(\hat{a}_0)$, where $\hat{a}_0$ is the zero order coefficient in the polynomial.

The estimation of the coherence, which is the square of the coherency, using the smoothed cross-periodogram ordinates was suggested and studied in [Tick 1967]. The idea of fitting a second-order polynomial is taken from the method of estimating the autospectrums.

The sign of the covariance, $e_{p_j p_k}$, is taken from the real part of the cross-periodogram at frequency $1/M$. The degree of freedom in the covariance matrice S is the same as the degree of freedom in the diagonal elements, which is provided by the spectral method.

4. EMPIRICAL STUDIES

The behaviour of the percentile estimation method given in the previous section was studied empirically using seven different stochastic processes. The simulated processes fall into three classes. For each process five percentiles, 10%, 25%, 50%, 75% and 90%, were estimated. The maximum allowed relative distance (ε) was 0.224. This implies that each percentile should have (on average) relative error less than 0.1. The confidence level was 0.9. 100 independent replications of each process were generated. All the processes were simulated under stationary conditions. The effect of initial transient was thus not studied.

The observations were collected in blocks of size 512 ($= n_b$). The initialization of the extended P^2-algorithm was done using the first block. The associated binary processes were divided into 128 ($= M$) subsequences. The degree of freedom in the estimated covariance matrice is 9 ($= f$). The termination criterion (inequality (5) in Section 2.1) was evaluated for the first time after the initialization. If the inequality is not satisfied, the sample size is doubled before the next evaluation of the termination criterion.

4.1 Simulated Processes

The first class of processes we analyzed were (stationary) turnaround time sequences, i.e. sums of the waiting time and the service time in M/M/1 queues. The second class were turnaround time sequences in $M/H_2/1$ queues. In a $M/H_2/1$ queue the service time distribution is a mixture of two exponential distributions. With probability π the service time is exponential with rate μ_1 and with probability $1 - \pi$ exponential with rate μ_2. The third class was the NEAR(1) processes introduced in [Lawrence and Lewis 1981]. The NEAR(1) process $\{X_n\}$ with parameters α and β is a Markovian sequence of exponentially distributed random variables with rate 1 and correlation structure $\mathrm{Corr}(X_n, X_{n+k}) = (\alpha\beta)^{|k|}$.

The seven processes we analyzed were:

1) M/M/1: $\lambda = 1\,,\ \mu = 2\,,\ \rho = 0.5$
2) M/M/1: $\lambda = 1\,,\ \mu = 1.25\,,\ \rho = 0.8$
3) M/M/1: $\lambda = 1\,,\ \mu = 1.111\ldots\,,\ \rho = 0.9$
4) M/H_2/1: $\lambda = 9\,,\ \mu_1 = 3.6\,,\ \mu_2 = 32.4\,,\ \pi = 0.1\,,\ \rho = 0.5$
5) M/H_2/1: $\lambda = 9\,,\ \mu_1 = 2.25\,,\ \mu_2 = 20.25\,,\ \pi = 0.1\,,\ \rho = 0.8$
6) NEAR(1): $\alpha = \beta = 0.95$
7) NEAR(1): $\alpha = \beta = 0.995$

4.2 Simulation Results

The results of the simulation studies are given in Figures 1–3 and Tables 1–2. Figure 1 contains the box-plots of the relative distances between the estimated and exact values. The run length characteristics are given in Table 1. The relative errors of each percentile are box-plotted in Figure 2. The interactions of the relative errors are visualized in Figure 3 using the biplot-technique. The standardized square roots of the mean square errors of each percentile are given in Table 2.

The general conclusion is that the achieved accuracy is excellent. In each replication the relative distance is less than the allowed maximum (ε). Typically the distance is only about 25% of the allowed maximum. The individual percentiles are also estimated accurately. Only in a few replications, the relative error of some percentile is greater than the allowed average of 10%. However, the price is quite high. The run lengths are long, but fortunately their variation is low.

The box-plots given in Figure 1 summarize the distributions of the relative distances between the estimated and exact values: $\left(\sum_{i=1}^{5}(1 - \hat{x}_{p_i}/x_{p_i})^2\right)^{1/2}$, where $\hat{x}_{p_i}$ is the estimated percentile and x_{p_i} is the exact percentile. The box-plot summaries are given by five percentiles and the extreme values. The ends of the box denote the 25^{th} and the 75^{th} percentile. The horizontal line inside the box denotes the median. The vertical lines reach the 10^{th} and the 90^{th} percentile. The minimum and the maximum are marked with separate horizontal lines.

The run length characteristics are given in Table 1. The given characteristics are geometric means, entropies and relative frequencies of different run lengths. Since the possible values of run lengths are powers of 2, the geometric mean G and the entropy E are the most suitable statistics to measure the location and the variation of the run lengths. The entropy is defined as $E = -\sum f_i \log f_i$, where the f_i's are the relative frequencies of different run lengths.

The characteristics of the relative errors are summarized in Figures 2-3 and Table 2. The relative error is defined as $(\hat{x}_{p_i} - x_{p_i})/x_{p_i}$. The marginal distributions of the relative errors of each

Figure 1: Box-plot of relative distances

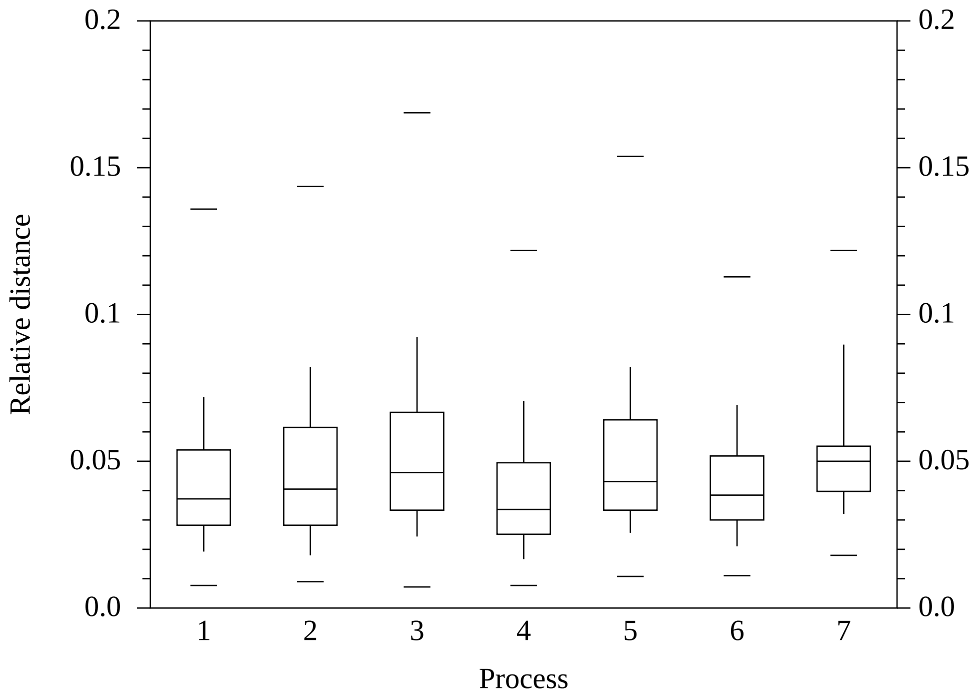


Table 1: Summary of Run Length Characteristics
Geometric Means, Entropies and Relative Frequencies

	Process						
	1	2	3	4	5	6	7
Geom. Mean	20738	153725	619180	125733	645474	67847	606438
Entropy	1.3922	1.3826	1.1767	1.2413	1.2300	1.1236	1.0366
Relative Frequencies							
8192	.07						
16384	.54					.01	
32768	.37					.09	
65536	.02	.12		.19		.74	
131072		.53		.69		.16	
262144		.35	.06	.11	.06		.04
524288			.64	.01	.58		.71
1048576			.30		.36		.25

Figure 2: Box-plot of relative distances

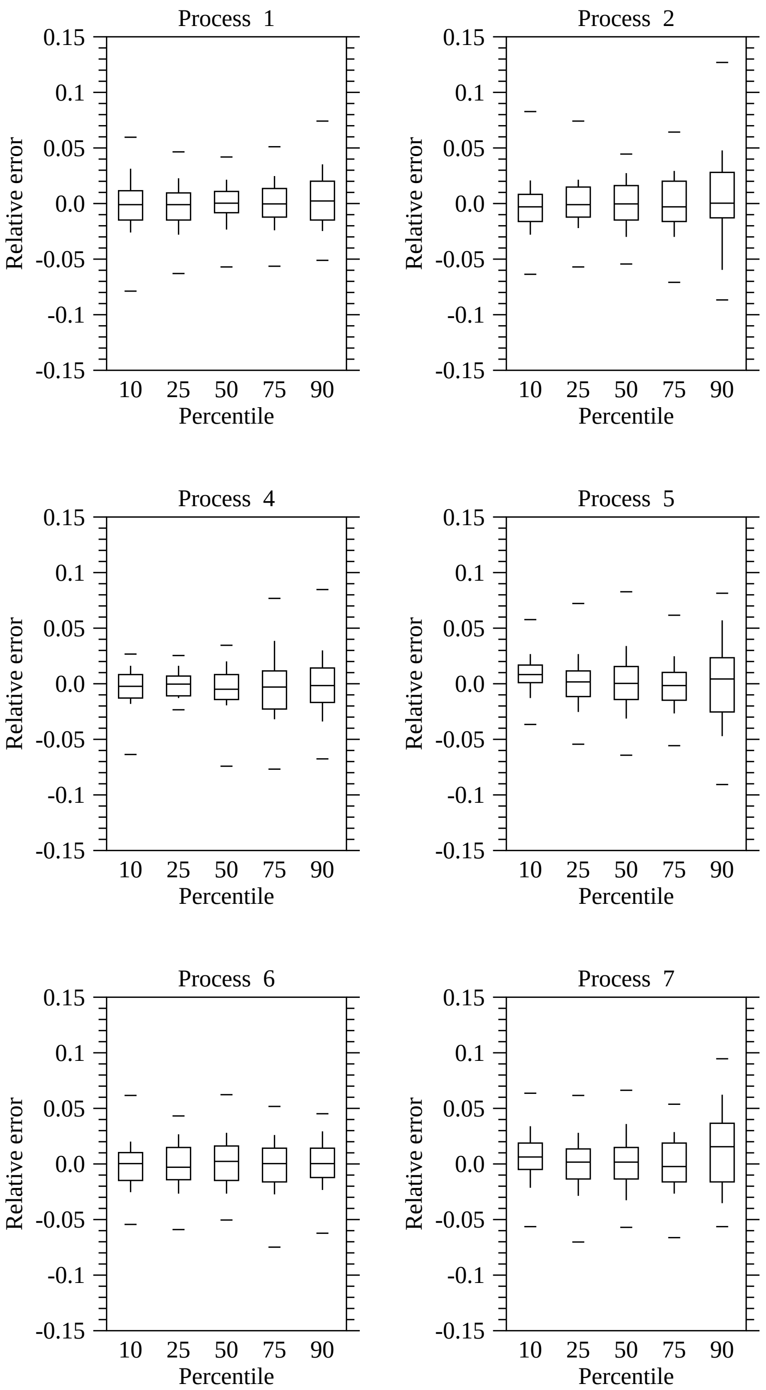


Table 2: Standardized Square Roots of Mean Square Errors

	Percentile				
Process	10 %	25 %	50 %	75 %	90 %
1	.0241	.0210	.0177	.0190	.0246
2	.0203	.0202	.0211	.0238	.0344
3	.0207	.0207	.0225	.0250	.0405
4	.0130	.0112	.0158	.0290	.0258
5	.0183	.0221	.0253	.0210	.0379
6	.0204	.0205	.0221	.0221	.0217
7	.0235	.0234	.0248	.0236	.0374

percentile are box-plotted in Figure 2. In Table 2 the standardized square roots of the mean square errors are given. The standardization is done by dividing the square root by the exact value of the percentile. This quantity measures the mean relative variation of the estimates from the true value.

In the given biplot (Figure 3) the covariance matrice of the relative errors is approximated by a matrice, whose rank is 2. The relative errors are represented as directed arrows starting from the origo. The length of the arrow approximates the standard deviation of the corresponding relative error. The cosine of the angle between two arrows represents the correlation coefficient. The technical details of the biplot-technique are given in [Gabriel 1971]. In Figure 3 the relative error of the median is directed to the east. This makes the comparison of different processes easier. The approximations are very accurate. In all the cases the accuracy is at least 98% in the least square sense.

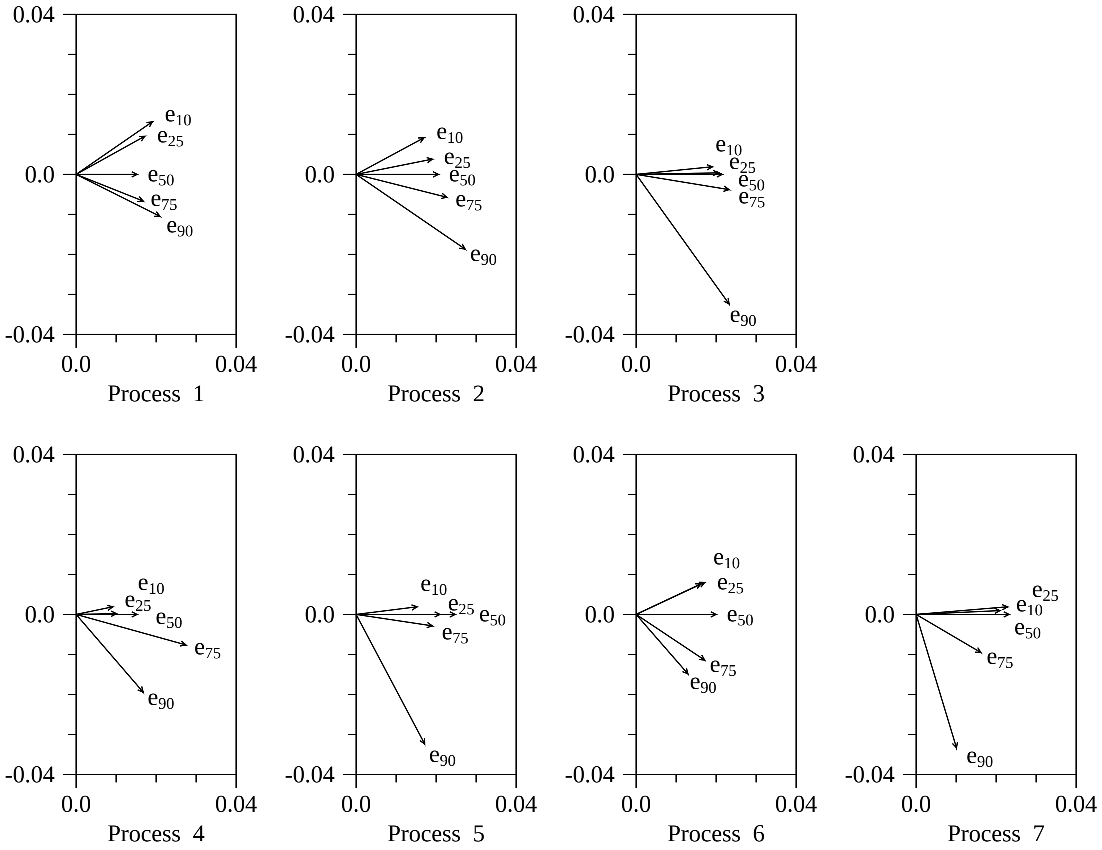
In M/M/1 and M/H_2/1 queues (Processes 1-5) the variance of the turnaround times increases as the traffic intensity (ρ) gets higher. The sequence also becomes more autocorrelated. Both these phenomena increase the difficulty of estimation. Their effects can easily be seen in Figure 1 and Table 1. The range of the relative distances gets wider, and the run lengths increase significantly. However, the variation of the run lengths does not increase.

In the NEAR(1) process (Processes 6-7) the same effects arise. However, the effects are only due to the increased autocorrelation, since the variance is independent of the parameters α and β.

The marginal properties given in Figure 2 and Table 2 indicate that the estimates of high percentiles have greater relative errors. Usually in independent samples, the median estimate is the most stable one of the percentiles. However, high autocorrelations increase the variation of median estimate so that it is equal to or greater than the variation of the lower percentiles.

Interesting effects can be found in Figure 3, where the interactions of the relative errors are biplotted. The increasing autocorrelation in the output sequence makes the relative errors of the

Figure 3: Biplot of relative errors



10^{th}, 25^{th} and 50^{th} percentile estimates almost linearly dependent. On the other hand, the relative error of the 90^{th} percentile estimate is almost independent of those three estimates.

The results reported above show that our method is able to produce estimates within the predefined accuracy. The achieved high level of accuracy is not unique. We have met the same phenomenon also in controling the accuracy of the mean, when the achieved accuracy is measured in terms of relative errors, but the method generates the confidence interval (in our method confidence region). We did not report the coverages of the estimated confidence regions, since the regions were used only as a tool in the termination decisions. The high level of correlations between the relative errors of the percentile estimates indicate that our approach to take the effects of covariances into account in the measure of accuracy is essential to the control of the total accuracy of the estimates.

The reported processes all have continuous marginal distribution. In queueing systems the waiting time is sometimes of interest. We have made some preliminary tests to see how the *extended* P^2-algorithm works in such cases. The none-zero probability of zero waiting time will cause slight errors in the estimates, whose expected value is zero. The reason is that the parabolic (or linear) correction of the estimate is unable to force a non-zero estimate to zero. Our opinion is that if a waiting time sequence is to be analyzed the estimation should be partitioned to estimate the probability of waiting and the conditional waiting time percentiles separately. In such a case our method will work better. In addition the analyst obtains more information about the sequence. The variance of the probability estimate can be calculated applying the spectral method to the binary sequence $\{I_j\}$, where $I_j = 1$, if the j^{th} waiting time is not zero. In order to reduce the amount of storage, this sequence should be batched dynamically.

Acknowledgement

The ideas and suggestions, which arose in discussions with Dr. Timo Alanko and Solveig Kurtén during the research and the preparation of this paper, have been very valuable and stimulating.

References

Gabriel, K.R. (1971) “The Biplot Graphic Display of Matrices with Application to Principal Component Analysis,” *Biometrika 58*, 453–467.

Heidelberger, P. and P.A.W. Lewis (1984), “Quantile Estimation in Dependent Sequences,” *Operations Research 32*, 185–209.

Heidelberger, P. and P.D. Welch (1981), “A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations,” *Communications of the ACM 24*, 233–245.

Jain, R. and I. Chlamtac (1985), "The P[2] Algorithm for Dynamic Calculations of Quantiles and Histograms without Storing Observations," *Communications of the ACM 28*, 1076–1085.

Lawrence, A.J. and P.A.W. Lewis (1981), "A New Autoregressive Time Series Model in Exponential Variables (NEAR(1))," *Advances in Applied Probability 13*, 826–845.

Raatikainen, K.E.E. (1986), "Simultaneous Estimation of Several Percentiles," *submitted to Simulation.*

Seila, A.F. (1982), "Estimation of Percentiles in Discrete Event Simulation," *Simulation 39*, 193–200.

Sen, P.K. (1972), "On the Bahadur Representation of Sample Quantiles for Sequences of ϕ-mixing Random Variables," *Journal of Multivariate Analysis 2*, 77–95.

Tick, L.J. (1967), "Estimation of Coherence," In *Spectral Analysis of Time Series* , B. Harris, Ed. Wiley, New York, pp 133–152.

APPENDIX

Extended P^2-Algorithm: To estimate the percentilies $\{\hat{q}_i\}_1^m$ and densities $\{\hat{f}(\hat{q}_i)\}_1^m$ of sequence $\{x_j\}_1^n$ at points $\{p_i\}_1^m$: $0 < p_1 < p_2 < \ldots.. < p_m < 1$

A Initialization: Sort the first n_b observations ($n_b \geq 2m+3$) and set

$c_1 \leftarrow 0$, $c_{2m+3} \leftarrow 1$	*increments of*
$c_{2i+1} \leftarrow p_i$, $i = 1, \ldots, m$	*desired positions*
$c_{2i} \leftarrow (c_{2i-1} + c_{2i+1})/2$, $i = 1, \ldots, m+1$	
$d_i \leftarrow 1 + (n_b - 1)c_i$, $i = 1, \ldots, 2m+3$	desired positions
$n_i \leftarrow [d_i]$, $i = 1, \ldots, 2m+3$	*actuals positions*
$q_i \leftarrow x_{(n_i)}$, $i = 1, \ldots, 2m+3$	*heights*

B For each subsequent observation x: $x \leftarrow x_j$, $j = n_b + 1, \ldots, n$ perform the following steps

1. Find sell k such that $q_k \leq x < q_{k+1}$
   If $x < q_1$ then $k \leftarrow 1$ and $q_1 \leftarrow x$
   If $x \geq q_{2m+3}$ then $k \leftarrow 2m+2$ and $q_{2m+3} \leftarrow x$
2. Increase actual positions of markers $k+1, \ldots, 2m+3$
   $n_i \leftarrow n_i + 1$, $i = k+1, \ldots, 2m+3$
3. Increase desired positions of all markers
   $d_i \leftarrow d_i + c_i$, $i = 1, \ldots, 2m+3$
4. Adjust heights and actual positions of markers $2, \ldots, 2m+2$, if necessary
   for $i = 2$ to $2m+2$ do

$d \leftarrow d_i - n_i$	*offset to desired position*
$dp \leftarrow n_{i+1} - n_i$	*offset to next position*
$dm \leftarrow n_{i-1} - n_i$	*offset to previous position*
$qp \leftarrow (q_{i+1} - q_i)/dp$	
$qm \leftarrow (q_{i-1} - q_i)/dm$	

if $d \geq 1$ and $dp > 1$ then
$qt \leftarrow q_i + ((1 - dm)qp + (dp - 1)qm)/(dp - dm)$
if $q_{i-1} < qt < q_{i+1}$ then $q_i \leftarrow qt$
else $q_i \leftarrow q_i + qp$
$n_i \leftarrow n_i + 1$
else if $d \leq -1$ and $dm < -1$ then
$qt \leftarrow q_i - ((1 + dp)qm - (dm + 1)qp)/(dp - dm)$
if $q_{i-1} < qt < q_{i+1}$ then $q_i \leftarrow qt$
else $q_i \leftarrow q_i - qm$
$n_i \leftarrow n_i - 1$
enddo

C Estimates: $\hat{q}_i = q_{1+2i}$, $i = 1, \ldots, m$

D Densities:

for $i = 1$ to m do
$j \leftarrow 2i + 1$
$\hat{f}_i \leftarrow \frac{n_{j+1} - n_{j-1}}{n} \left(\frac{n_j - n_{j-1}}{n_{j+1} - n_j} (q_{j+1} - q_j) + \frac{n_{j+1} - n_j}{n_j - n_{j-1}} (q_j - q_{j-1}) \right)$
enddo