Quantile estimation in regenerative simulation
Coppens, G.W.J.; van Dongen, M.P.F.M.; Kleijnen, J.P.C.

Publication date: 1975

Citation for published version (APA):
REEKS "TER DISCUSSIE"

FACULTEIT DER ECONOMISCHE WETENSCHAPPEN
QUANTILE ESTIMATION IN REGENERATIVE SIMULATION:
A CASE STUDY.

G. Coppens, M.v. Dongen, J. Kleijnen

FACULTEIT DER ECONOMISCHE WETENSCHAPPALEN
QUANTILE ESTIMATION IN REGENERATIVE SIMULATION:
A CASE STUDY

BY

G.W.J. Coppens, M.P.F.M. van Dongen and J.P.C. Kleijnen

Department of Business and Economics
Katholieke Hogeschool, Tilburg, Netherlands

July 1976
ABSTRACT: We model key-punching in a computer center as a queuing simulation with 2 servers (typists) and 3 priority classes (small, medium, large jobs). The 90% quantile of queuing time is estimated for different borderlines between the 3 job classes. Confidence intervals for the quantiles are based on the regenerative properties of the simulation, as derived by Iglehart (1974). They utilize the asymptotic normality of the estimated quantile, and a rather complicated expression for its variance. Numerical results are given for the quantiles (and averages) of the queuing times in each job class, for several borderlines between the 3 job classes. The effects of simulation runlength on the confidence intervals were also examined. The effects of varying job-class borderlines were tentatively modeled by a regression model.

1. INTRODUCTION
Our interest in the topic of this paper was generated by the following practical problem. The computer center of our university provides key-punching services to all its users (students, faculty members, administration). There are 2 typists (servers) available. To provide a "reasonable" turnaround time the following priorities are established. "Small" jobs (S-jobs) have priority over "medium" (M) jobs which in turn are key-punched before "large" (L) jobs. Within each of the 3 priority classes a first-come-first-served rule applies. The priority system is not preemptive, i.e., a job currently processed is not interrupted by a higher-priority job, but is first finished. The computer center's management wanted to know how to choose the borderline X between small and medium jobs, and the limit Y between medium and large jobs. (Symbols are defined in table 1.) Classification of jobs is possible indeed, since service time (i.e. key-punching time) for a job can be accurately enough predicted from the handwritten code sheets, which are to be key-punched. The criterion for choosing X and Y is the queuing time, i.e. the waiting time excluding key-punching (servicing) itself. The interarrival times (in minutes) were found to be exponentially distributed with parameter \( \lambda = 0.033 \). The service times are also exponential with parameter \( \mu = 0.021 \), and a minimum value of 9.5625 minutes (and a maximum of 900 minutes or 15 hours, whereas theoretically an exponential distribution can yield infinite values). Currently the computer center classified the jobs as small, medium or large using \( X = 30, Y = 180 \). During our investigation it soon turned out that management...
was not so much interested in the average queuing time, but it wanted to help "as many people as fast as possible". It was agreed that we could formulate this criterion as the 90% quantile, i.e., if $x$ denotes the queuing time, then the quantile $Q_{0.90}$ is such that (underlining stochastic variables)

\[(1.1) \quad P(x \leq Q_{0.90}) = 0.90\]

So there is only a 10% chance that customers have to wait longer than $Q_{0.90}$. Note that if $x$ is discrete instead of continuous, then the definition of $Q_{0.90}$ is a bit more complicated; see Kleijnen (1975, p. 478).

From our problem formulation it follows that we formulated our system as a queuing system with 2 service stations (typists) and 3 priority classes, which are based on the lengths of the service times. This problem has not yet been solved analytically, so that we use simulation. We want to estimate the quantile with known statistical accuracy, i.e., we want to compute a confidence interval for $Q_{0.90}$. The next step is to compute the effects of the control limits $X$ and $Y$ (priority class-limits) on $Q_{0.90}$ in order to minimize $Q_{0.90}$. To compute confidence intervals we analyze the simulation using its regenerative property, i.e., if the system becomes empty (both servers idle), then a new history starts independently of the past simulated history; see Iglehart (1974), Kleijnen (1975, p. 46) and fig. 1. Observe that the utilization or traffic intensity is $\lambda/2\mu = 0.79 < 1$, so that the queuing system is stable.

2. Iglehart's Quantile Estimation

Iglehart (1974) derived the following approach for the confidence intervals of a quantile in a simulation with the regenerative property. Let us simulate $n$ independent regenerated cycles, comprising a total of $\bar{B}_n$ customers. We can count the number of customers having a queuing time $x$ smaller than a constant, say $x$. This number of customers divided by the total number of customers leads us to the empirical distribution function, say $\hat{F}$. We should, however, distinguish here between a discrete variable $x$ and a continuous $x$. If, for instance, $x$ can only assume the values 2, 3, 4, then indeed we can count how many customers show the value 2, how many show the value 3, etc., and divide these numbers of customers by their total $\bar{B}_n$. If $x$ is a continuous variable, then we can sort all $\bar{B}_n$ customers in ascending order, and the empirical distribution $\hat{F}$ makes a jump of magnitude $1/\bar{B}_n$ at each observed value of $x_i$ ($i = 1, \ldots, \bar{B}_n$). See also Kleijnen (1975, p. 478).

Since sorting is time-consuming Iglehart proposes to discretize the continuous variable by dividing its range into $N$ classes (with fixed, but possibly different, lengths). Denoting the number of customers in class $j$ by $\bar{w}_j$, $\hat{F}$ shows a jump of magnitude $\bar{w}_j/\bar{B}_n$ at class $j$; see fig. 2.

A problem in discretization is that fig. 2 yields only the class into which the quantile falls, not a unique value. Therefore we interpolate linearly as shown by the dashed lines $\hat{F}'$ in fig. 3, where as an example we assume that $0 \leq x \leq 5$, and $N = 5$. Denote the resulting quantile by $\hat{Q}$.

In order to compute a confidence interval for the estimated $\hat{Q}_{0.90}$ we use the regenerative characteristics of the simulation. (To simplify our notation we may drop the lower index .90 for $\hat{Q}$.) Denote the number of customers within cycle $k$ having a queuing time not exceeding $x$, by $\bar{Y}_k(x)$. If there are $\bar{B}_k$ custo-
mbers in that cycle, then the expected value of \( Y_k \) is \( \mu_k \). Then introduce

\[
(2.1) \quad z_k(x) = Y_k(x) - F(x) \mu_k
\]

Obviously the \( z_k(x) \) are identically and independently distributed since the regenerative property makes each cycle \( k \) independent of each other cycle. We know that the expected value of \( z \) is zero. Denote its variance briefly by \( \sigma^2 \), i.e.,

\[
(2.2) \quad \sigma^2(x) = \text{var}(z_k(x))
\]

Under certain technical restrictions - see Iglehart (1975, p.13) - the estimated quantile \( \hat{Q} \) is asymptotically normally distributed with variance

\[
(2.3) \quad \text{var}(\hat{Q}) = \frac{E(\mu)}{E(\mu) \cdot \hat{Q}'}^2 \cdot \sigma^2(\hat{Q}) \cdot n
\]

where the various factors will be discussed next.

(i) the expected value \( E(\mu) \) can obviously be estimated by

\[
(2.4) \quad \mu = \frac{1}{n} \sum_{k=1}^{n} \mu_k
\]

Note that as the cycles comprise more customers, the variance tends to decrease, as we expect intuitively.

(ii) \( \hat{Q}'(\hat{Q}) \) denotes the derivative of \( \hat{Q} \), evaluated at \( x = \hat{Q} \). This derivative is estimated by the slope \( \hat{F}_n \) shown in fig. 4, i.e.

\[
(2.5) \quad \hat{F}_n = \frac{w}{H-L}
\]

where \( w \) denotes the number of customers falling in the relevant class, and \( L \) and \( H \) are the lower and higher limits of that class. (In Iglehart, 1974, p. 15, \( H-L=1 \). We do not explicitly show the class-index of \( w \).)

(iii) \( \sigma^2(\hat{Q}) \) requires a more complicated procedure. Iglehart (1974, p.15) derived that in a point \( x \) (including \( x=\hat{Q} \)) we have

\[
(2.6) \quad \sigma^2(x) = \left[ \frac{1}{n} \sum_{k=1}^{n} (x - \hat{Q})^2 \right] \cdot \text{var}(Y_k(x)) - \{ \hat{Y}(x) \}^2 - \{ \hat{Y}(x-1) \}^2
\]

where \( |x| \) denotes the integer part of \( x \) (entier \( x \)). We can estimate \( \sigma^2(x) \) for \( x = |x| \) or \( |x| + 1 \) occurring in eq. (2.6), as follows. Using eqs. (2.2) and (2.1) we have

\[
(2.7) \quad \sigma^2(x) = \text{var}(z_k(x)) = \text{var}(Y_k(x) - F(x) \mu_k),
\]

and

\[
(2.8) \quad \text{var}(Y_k(x)) = \frac{1}{n-1} \sum_{k=1}^{n} (Y_k(x) - \bar{Y}(x))^2
\]

\[
(2.9) \quad \text{var}(\mu_k) = \frac{1}{n-1} \sum_{k=1}^{n} (\mu_k - \bar{\mu})^2
\]

\[
(2.10) \quad \text{cov}(Y_k, \mu_k) = \frac{1}{n-1} \sum_{k=1}^{n} (Y_k(x) - \bar{Y}(x)) \cdot (\mu_k - \bar{\mu})
\]

where we use the traditional estimators

\[
(2.11) \quad F(x) = \frac{1}{n} \sum_{k=1}^{n} Y_k(x)
\]

In (2.6) we use (2.7) through (2.11), once for \( x = |x| \) \( = \hat{Q} \) and once for \( x = |x| + 1 \) \( = \hat{Q} + 1 \). Moreover we need the following estimator in (2.6):

\[
(2.12) \quad E\{Z(|x|) \} \cdot E\{Z(|x|+1) \} \cdot \sum_{k=1}^{n} (Y_k(|x|) - \bar{Y}(|x|)) \cdot \mu_k
\]

See also Iglehart (1974, p. 15-16).

Note that in the above formulas we need to know \( Y_k(|x|) \) and \( Y_k(|x|+1) \). However, \( \hat{Q} \) is estimated after we have simulated \( n \) cycles. Hence these \( Y_k \)'s can be computed only afterwards, so that in the mean time we have to store all \( \hat{Q}_n \) individual queuing times. Observe further that the selection of the classes into which the individual queuing times are placed (discretization), is performed heuristically as follows. Divide the whole range (from 0 to maximum observed queuing time) into 20 classes of equal width, and compute a first estimate of \( \hat{Q} \) as shown in fig. 3. A second, refined estimate is based on a subdivision of the relevant class. Several methods for subdividing this class were tried; they all resulted in approximately the same value for \( \hat{Q} \). One such me-
3. NUMERICAL RESULTS FOR VARYING CONTROL LIMITS X AND Y

Remember that X and Y determine the borderlines between small and medium, and between medium and large jobs respectively. For each X,Y combination and each job class we computed the following statistics based on 500 cycles: number of customers per job class\(^1\), average service (key-punching) time\(^2\), average queuing (excluding service) time, 90% quantile \(\hat{Q}\), its 90% confidence interval, and the length of the confidence interval expressed as a percentage of \(\hat{Q}\). (To save space these tables are not reproduced here; they can be obtained from the authors\(^3\).) The relative lengths of the confidence intervals of \(\hat{Q}\) were between 2.89% and 5.83% for S-jobs, 5.24% and 25.39% for M-jobs, 9.28% and 83.51% for L-jobs.

To see the effects of increasing the number of cycles (n), we took n=200, 300, 400, 500, 1000, 1500, 2000 and computed the average queuing time per job class for 2 \(X,Y\) combinations. Especially L-jobs did not stabilize as smoothly as we had hoped for. See also fig. 5. Besides the average queuing time we computed the relative lengths of the confidence intervals for \(\hat{Q}\); see fig. 6. S- and M-jobs showed the expected decreasing lengths of the confidence intervals. We found that one particular cycle, viz. cycle 479, gave results that disturbed the overall picture.

Next we looked at the effects of the X and Y borderlines on the average queuing time and the 90% quantile for each of the 3 job classes. We also computed the average queuing time over all 3 job classes (weighted with the number of customers per class). The resulting pictures did not show a clear pattern of possible influences of X and Y; for illustration purposes 1 of the 7 pictures is shown in fig. 7. This visual analysis was supplemented by the following formal analysis based on a regression (meta)model.

\[
(3.1) \quad y = y_0 + y_1 X + y_2 Y + y_11 X^2 + y_12 XY + y_22 Y^2 + e
\]

Here \(y\) denotes the response variable, e.g. the 90% quantile for S-jobs. In total we studied 7 response variables, viz. 3 quantiles and 3 averages per job class and 1 overall average. X and Y denote the job-class limits. We studied 19 combinations of X and Y; see fig. 7. The inaccuracy of the model is denoted by \(e\). The 6 regression parameters \(y\) were estimated by the least squares method. The analysis also yielded the following statistics.

1. The t-values of the \(\hat{y}\)'s used to test whether \(y\) deviates significantly from zero.
2. An F-value to test if all \(y\)'s (excluding \(y_0\)), taken together, are zero, i.e. X and Y have no effect on \(y\).
3. The correlation coefficient \(R^2\).

An example of our results is shown in eq. (3.2), which specifies (3.1) for the 90% quantile of small jobs, say \(Q_{90}^{S}\).

\[
(3.2) \quad Q_{90}^{S} = 57.24 - 0.04336X - 0.01005Y + 0.007182X^2 + 0.001102XY - 0.0002519Y^2
\]

with \(R^2=0.999\), F=17.12 and t-values 1.458, 0.066, 0.023, 0.810, 0.403, 0.208 respectively. Summarizing all 7 regression equations, we saw that all \(R^2\)'s were high, namely between 0.863 and 0.999. Nevertheless it turned out that all t-tests were insignificant at \(\alpha = 0.05\), i.e. each \(y\) may be assumed to be zero. All F-tests except for the overall queuing time and the quantile
for L-jobs, were significant, i.e., X and Y do have effects. Whether the postulated model (3.1) is a good model, was tested using an F-test for lack-of-fit; see Kleijnen (1975, p. 367). This test was only applied for the quantiles of the S-jobs, and was found to be significant, i.e., the model (3.1) is no adequate model\(^4\). Observe that the tests assume independent, normally distributed variables with constant variances. We did not investigate how seriously deviations from these assumptions affect the tests in our study. (For guidelines see Kleijnen, 1975, p. 718-725.) The conclusions based on the statistical tests do not contradict the intuitive, visual interpretation of the experimental results.

Summarizing, though the regression model (3.1) yielded high R\(^2\) values for each response variable \(y\), the model is nevertheless inadequate as the lack-of-fit F-tests showed. The X and Y limits do have effects (see the F-tests for the joint \(y\)), but more research is needed to find the correct form of their effects\(^5\). Therefore eq. (3.1) was not used to find the X,Y combination which would result in the lowest value for the 90\% quantile (or the average).

4. CONCLUSION

We hope that our paper will make some simulation practitioners aware of the existence and nature of Iglehart's procedure for quantile estimation in simulation experiments with the regenerative property. Moreover, we showed some numerical results for varying lengths of the simulation runs. The effects of the control limits X and Y were modeled by a regression equation that was tested and, unfortunately, rejected.

FOOTNOTES

1) Using the same random numbers, an increase of X from 30 to 35 for fixed Y(=150) decreased the number of S-customers from 7074 to 6325. An increase of the number of S-customers was expected, until we realized that for a fixed number of cycles the total number of customers changes when changing X.

2) Could be checked against input parameters of service distribution.

3) A sample from these tables is as follows. For X=30, Y=150 we obtained 7074 S-customers with average service time 18.98 (minutes), average queuing time 29.67, maximum queuing time 292.34, 90\% quantile 66.64, 90\% confidence interval from 64.45 to 68.82. For X=35, Y=150 these numbers became: 6325 (customers), 21.24 (service), 28.97 (queuing), 211.02 (maximum), 64.87 (quantile), 62.31 to 67.42 (interval).

4) The residual sum of squares based on 19 (X,Y) combinations was compared with the "pure" experimental error estimated at each (X,Y) combination from 500 independent cycles using eq. (2.3). Of the 19 combinations 17 yielded significant F-values.

5) After this preliminary investigation was closed, A. van Reeken (manager computer center, Katholieke Hogeschool) suggested the following transformations.

(i) In eq. (3.1) replace \(Y\) by \((Y-X)\), i.e. the increment of \(Y\) over \(X\). Then we still obtain a second degree polynomial in \(X\) and \(Y\) like (3.1), but with different coefficients, viz.

\[
Y = \gamma_0 + (\gamma_1 - \gamma_2)X + \gamma_2 Y + (\gamma_{12} - 2\gamma_{22})XY + (\gamma_{11} + \gamma_{22} - \gamma_{12})X^2 + \gamma_{22}Y^2 + \epsilon
\]

(ii) Replace \(X\) by the probability of a small job, i.e., a service time smaller than \(X\), or

\[
P_1 = \int \mu \exp(-\mu x) dx \quad (\mu=0.021)
\]

\[
a = 9.5625
\]

and replace \(Y\) by the probability of a service time between \(X\) and \(Y\), i.e.

\[
P_2 = \int \mu \exp(-\mu x) dx
\]

\[
Y
\]
It is: G. L'Ant. U.L. Simulat.iny stable stochas-
tic systems VI: duantile estimation. Pecl~nical report no. 86-75, Contr.rol Ana-
lysis Corporation, 800 Welch Road, Palo Alto, California, Dec. 1974. (Also publish-
ed in: Journal Association Computing Ma-
cyhine, 23, no. 2, April 1976, p. 347-
360.)

KLEIJNEN, J.P.C. Statistical techniques in
simulation (in two parts). Marcel Dekker

NOTE
This paper is based on a term project per-
formed by the first two authors, gradu-
ate students, under the supervision of the
third author, senior research associate,
at the Katholikke Hogeschool Tilburg. The
program was written in ALGOL-68 and run on
an ICL-1903 computer. To generate the data
listed in footnote 3, for all 19 X,Y com-
binations, 3 job classes and 500 cycles, we
needed 1 hour and 50 minutes of computer
time. The whole project required approxima-
tely 13 man-weeks.

\[ a_k = \text{number of customers in cycle } k \]
\[ \theta_n = \text{total number of customers in } n \text{ cy-
cles} \]
\[ \gamma = \text{regression parameter} \]
\[ e = \text{disturbance (noise, error) in re-
gression equation} \]
\[ F(x) = P(x < x) = \text{distribution function of } x \]
\[ \hat{F}(x) = \text{linear interpolation of } F(x) \]
\[ f = \text{estimated slope of } \hat{F} \]
\[ i = \text{index of customer } (i=1,\ldots,\theta_n) \]
\[ j = \text{index of class for discretized} \]
\[ \text{queuing time } (j=1,\ldots,N) \]
\[ k = \text{index of cycle } (k=1,\ldots,n) \]
\[ N = \text{number of classes } (N=1,\ldots,n) \]
\[ n = \text{number of cycles} \]
\[ Q_{90} = 90\% \text{ quantile} \]
\[ \hat{q} = \text{quantile based on } \hat{F} \]
\[ \sigma^2(x) = \text{variance based on discretization; see eq. (2.6)} \]
\[ \hat{\sigma}^2 = \text{variance based on discretization; see eq. (2.6)} \]
\[ w = \text{number of customers in class } j \]
\[ X = \text{borderline between small and me-
dium jobs} \]
\[ x = \text{queuing (=waiting excluding ser-
vicing) time} \]
\[ \lceil x \rceil = \text{integer part of } x \]
\[ Y = \text{borderline between medium and} \]
\[ \text{large jobs} \]
\[ Y_k(x) = \text{number of customers with } x < \chi \text{ in} \]
\[ \text{cycle } k \]
\[ y = \text{response variable in regression} \]
\[ \text{equation} \]
Figure 1: Regenerative queuing simulation
Figure 2: Empirical distribution function $\hat{F}$
Figure 3. Linear interpolation for quantile

Figure 4. Estimating $\hat{F}'(\tilde{Q})$
Figure 5. AVERAGE QUEUING TIME AND RUNLENGTH

SMALL JOBS

MEDIUM JOBS

LARGE JOBS

(X = 30  
Y = 180)

(X = 40  
Y = 150)

(X = 30  
Y = 180)

(X = 40  
Y = 150)

(X = 30  
Y = 180)

(X = 40  
Y = 150)
Figure 6. RELATIVE QUANTILE CONFIDENCE INTERVAL, AND RUNLENGTH
Figure 7. ESTIMATED QUANTILE OF SMALL JOBS AND X, Y COMBINATION

Y

220

210

200

190

180

170

160

150

140

X

20

25

30

35

40

45

50

75.54

57.18  59.34  63.84

71.26

67.99

64.24  63.21  61.10  70.15  73.05

68.57

72.23

58.09  66.64  64.87  68.69  71.31

69.22