European Journal of Operational Research 62 (1992) 372–379
North-Holland

Theory and Methodology

Sampling for quality inspection and correction: AOQL performance criteria

J.P.C. Kleijnen, J. Kriens, M.C.H.M. Lafleur and J.H.F. Pardoel
School of Business and Economics, Tilburg University, Tilburg, Netherlands

Received May 1990; revised February 1991

Abstract: By definition, an Average Outgoing Quality Limit (AOQL) sampling plan leads to inspection of the whole population if the sample shows a number of defective items k exceeding an acceptance number k_0. The literature shows how this constant k_0 and other related parameters can be chosen such that the expected value of $\tilde{p}$, the fraction of defectives after inspection and possible correction, does not exceed a prespecified constant $\tilde{p}_m$. This paper studies several other criteria that are ignored in the literature. It is based on an extensive Monte Carlo simulation. Its main conclusion is that AOQL sampling is useful in practice, including applications in auditing. Yet the probability that the average yearly outgoing fraction $\bar{\tilde{p}}$ exceeds the given constant $\tilde{p}_m$ can be sizable, if the original before-sampling fraction p exceeds $\tilde{p}_m$ 'mildly'. The paper further investigates the effects of splitting the yearly population into subpopulations and the effects of underestimating the original fraction.

Keywords: Percentage defective; quality control; Monte Carlo simulation; auditing

1. Introduction: AOQL

AOQL sampling plans were originally designed for quality control in industry. Nowadays they are also applied in auditing, which inspired this paper. This contribution evaluates these sampling schemes, using several criteria neglected in the literature, especially the probability of quality violations in the short run, say a year.

AOQL sampling plans were introduced by Dodge and Romig around 1930; see Dodge and Romig (1959). These plans are discussed in the monographs by Hald (1981, pp. 116–124) and Schilling (1982, pp. 372–399). Their practical application to auditing is studied by Kriens and Veenstra (1985). Nowadays, further interest in quality control is stimulated by the Japanese management philosophy; see Cross (1984) and Wurnik (1984). (Some of these references are the result of an extensive computerized literature search.)

The *goal* of AOQL sampling is to guarantee a minimum quality after inspection; this quality is expressed as a maximum $\tilde{p}_m$ for the expected value of the fraction of defectives in the population. For application in auditing Kriens and Veenstra (1985) split the yearly population into a number of subpopulations. The 'original' quality of the yearly population – before sampling and correction – is quantified by p, the fraction of 'defective items' in the yearly population; also see the symbol list in Table 1. The yearly population

Correspondence to: J.P.C. Kleijnen, School of Business and Economics, Katholieke Universiteit Brabant (Tilburg University), P.O. Box 90153, 5000 LE Tilburg, Netherlands.

Table 1
Symbol list (in alphabetical order; random variables are underscored)

J	= Number of simulated years in simulation
k	= Number of defective items in sample
k_0	= Acceptance number
K	= Number of defective items in subpopulation
n	= Sample size
N_s	= Subpopulation size in subperiod s ($s = 1,\ldots,S$)
NY	= Estimated number of items per year
p	= Original (before sampling) fraction of defectivenes in yearly population
$\hat{p}$	= Estimate of p
$\tilde{p}$	= Fraction of defectives after inspection and possible correction
$\tilde{p}_s$	= Outgoing fraction of defectives in period s ($s = 1,\ldots,S$)
$\bar{\tilde{p}}$	= Yearly outgoing fraction of defectives: $\bar{\tilde{p}} = \Sigma_{s=1}^{S} \tilde{p}_s(N_s/\text{NY})$
$\tilde{p}_m$	= Prespecified constant (not to be exceeded by the expected value of $\tilde{p}$)
q	= Probability of quality violation: $P[\bar{\tilde{p}} > \tilde{p}_m]$
$\hat{q}_j$	= Estimate of q after j replications
R^2	= Measure of fit
S	= Number of subpopulations per year
t	= time
$\underline{x}$	= binomial variable with parameters J and q.

is not known until the end of the year; hence it must be estimated. The estimated Number of items (correct plus defect) per Year is NY; for example, a company is expected to produce NY cars per year; in auditing, accounts are sampled and NY is measured in dollars per year. Consequently, after inspection and correction, the quality limit means that the expected value of the remaining fraction of defectives $\tilde{p}$ remains under a maximum value $\tilde{p}_m$, the so-called Average Outgoing Quality Limit.

The sampling scheme has the following steps (also see Table 2 later on).

(i) At the beginning of the year the accountants estimate the yearly population size NY. They also decide on the number of subpopulations S; for example, $S = 52$ corresponds to weeks. At the end of period s the size of the subpopulation turns out to be N_s ($s = 1,\ldots,S$). The choice of S depends on the organization.

(ii) From each realized subpopulation, a sample of size n is taken (n depends on several parameters).

(iii) Per sample the number of defective items $\underline{k}$ is determined by inspection. Obviously $\underline{k}$ is random, and the integer values k satisfy: $0 \le k \le n$.

(iv) If and only if k exceeds a critical constant k_0 (which varies with n), the whole subpopulation is inspected and, by assumption, all defective items in the *subpopulation* are corrected perfectly. (In auditing, defectives are errors that are often removed by corrective actions; Hald, 1981, pp. 311–312, discusses imperfect inspection and correction of items.) If, however, $k \le k_0$, then only the defective items found in the *sample* are corrected. So after this sampling, the quality of the subpopulation is improved, unless no defectives at all were found ($k = 0$).

Denote the outgoing fraction of defectives in period s by $\underline{\tilde{p}}_s$. Then the outgoing fraction of defectives in the yearly population is

$$\underline{\bar{\tilde{p}}} = \sum_{s=1}^{S} \underline{\tilde{p}}_s \left(\underline{N}_s / \sum_{s=1}^{S} \underline{N}_s \right). \tag{1.1}$$

Note that $\bar{\tilde{p}}$ reduces to $\tilde{p}$ if there are no subpopulations. The auditor wishes the average outgoing fraction $\bar{\tilde{p}}$ not to exceed the limit of defectiveness, $\tilde{p}_m$. So, given a correct selection of the sampling plan's parameters n and k_0, the yearly outgoing fraction $\bar{\tilde{p}}$ should satisfy the condition $E[\bar{\tilde{p}}] \le \tilde{p}_m$. Obviously, if the original (before sampling) fraction was very good already (say, $p = 0$), then $E[\bar{\tilde{p}}] \le \tilde{p}_m$. If this quality was very bad ($p \gg \tilde{p}_m$), then the sampling plan implies that sampling is (nearly) always followed by inspection and correction of the whole subpopulation, so $E(\underline{\bar{\tilde{p}}}) \ll \tilde{p}_m$. This gives Figure 1 where

Table 2
Sample size n and acceptance number k_0, given before-sampling fraction p, subpopulation size N_s, and defectiveness limit $\tilde{p}_m$; here $\tilde{p}_m = 1\%$

Subpopulation size N_s	Before-sampling fraction p								
	0–0.02		···	0.21–0.40		···	0.81–1.00		
N_s	n	k_0		n	k_0		n	k_0	
1–25	All	0		All	0		All	0	
26–50	22	0		22	0		22	0	
⋮									
801–1000	35	0		80	1		120	2	
1001–2000	36	0		80	1		180	3	
⋮									
20001–50000	85	1		255	4		990	15	
50001–100000	85	1		255	4		1520	22	

p^* is the 'least favorable' value of p; this figure assumes that there are no subpopulations.

Next, we consider the sampling plan's parameters. Because sampling is without replacement, $\underline{k}$ follows the hypergeometric distribution with parameters n, p and N_s. The literature proves that the critical constant k_0 and the sample size n can be computed such that the condition $E[\underline{\tilde{p}}] \leq \tilde{p}_m$ holds; moreover the expected costs can be minimized if p is known; we shall return to this issue. Unfortunately, the original tables in Dodge and Romig (1959) contain some inaccuracies; see Hald (1981, p. 124) and Van Batenburg, Kriens and Veenstra (1988). Therefore we use our own tables. Table 2 gives an example of a part of such a table. Tables for very small $\tilde{p}_m$-values are given in Cross (1988), while Wurnik (1984) gives nomograms for $k_0 = 0$.

In *practice* the before-sampling fraction p is unknown. In some applications the right most column is used; in other applications the left most column is taken. Dodge recommended use of the right most column for at least two reasons: (i) the sample sizes are larger for most N_s and hence, more reliable estimates of p are generated faster, and (ii) these sampling plans are generally more discriminating; also see Schilling (1982, p. 375). Our study, however, originates from questions raised by Dutch auditors. They always try to create situations with values of p as small as possible. Only if the auditors expect p to have a small value, will they apply statistical sampling procedures; therefore, if the AOQL procedure is used, they take the left most columns of tables like Table 2.

Even if p is estimated wrongly, the quality constraint $E[\underline{\tilde{p}}] \leq \tilde{p}_m$ is satisfied; the expected costs, however, may increase. Moreover, practitioners usually *conjecture* that the probability of excessive defectiveness is negligible. Figure 1 shows that if the original fractions p were always least favorable ($p = p^*$) and $\underline{\tilde{p}}$ were distributed symmetrically, then the probability of a quality violation would be 50%: $P(\underline{\tilde{p}} > \tilde{p}_m) = 0.50$. Practitioners presume that actually the probability for the average fraction is nearly zero: $P(\bar{\underline{\tilde{p}}} > \tilde{p}_m) \approx 0.00$. This conjecture is the main focus of our simulation. (Hald, 1981, p. 310, gives analytical approximations for this probability.)

It hardly takes more computer simulation time to estimate how bad the value $\bar{\underline{\tilde{p}}}$ is *if* the constraint $\bar{\underline{\tilde{p}}} < \tilde{p}_m$ is violated. Therefore we also estimate the following conditional expectation:

$$E\left[\bar{\underline{\tilde{p}}} - \tilde{p}_m \mid \bar{\underline{\tilde{p}}} > \tilde{p}_m\right]. \tag{1.2}$$

The next sections will show that this paper has the following contributions:

(i) It quantifies the effects of splitting the estimated yearly population (NY) into S *subpopulations*. A higher S leads to a lower expected constraint violation $E(\bar{\underline{\tilde{p}}} - \tilde{p}_m \mid \bar{\underline{\tilde{p}}} > \tilde{p}_m)$ (as Figure 4 will show).

(ii) It quantifies not only the *average* yearly outgoing fraction $\bar{\underline{\tilde{p}}}$ (Figure 2) but also the *probability* of a constraint violation $P(\bar{\underline{\tilde{p}}} > \tilde{p}_m)$ (Figure 3). That probability may be as high as 40%, which is certainly not negligible!

(iii) It estimates the effects of *underestimating* the before-sampling fraction p if practitioners

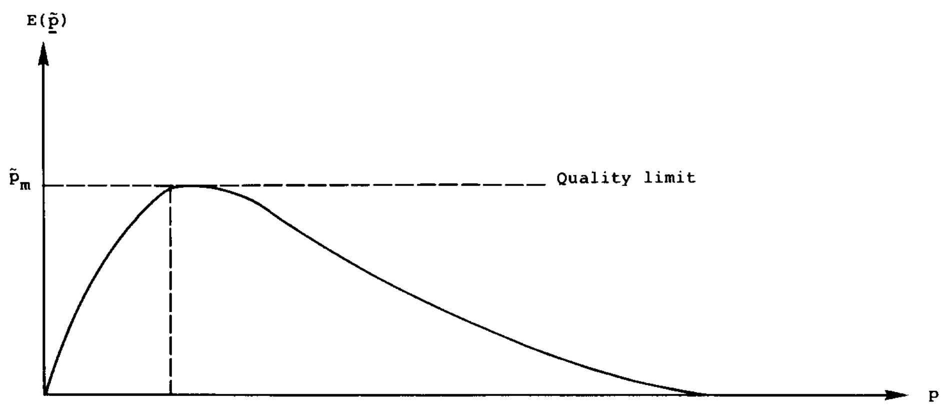

Figure 1. Expected fraction of defectives after sampling $E(\underline{\tilde{p}})$ versus fraction before sampling p (different scales on different axes)

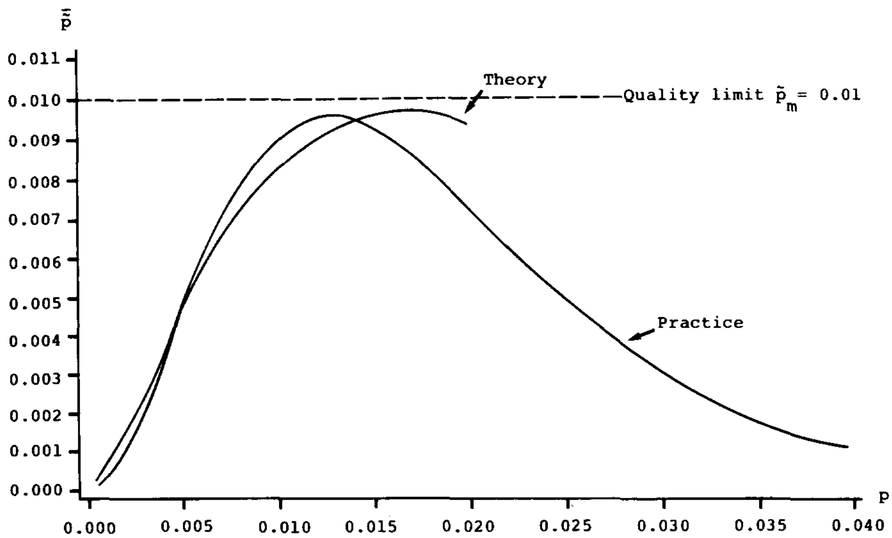

Figure 2. Average yearly outgoing quality $\bar{\bar{p}}$ versus fraction of defectives before sampling (yearly population NY = 1000000; subperiods $S = 52$)

use only the left most columns of tables like Table 2. This practice results in higher *costs* (Figure 5) while the *probability* of a constraint violation may nevertheless increase (Figure 3).

(iv) It gives more insight into AOQL plans. For example, higher variability in the before-sampling fraction p (over subpopulations) gives additional protection (see Section 2). Estimation of p

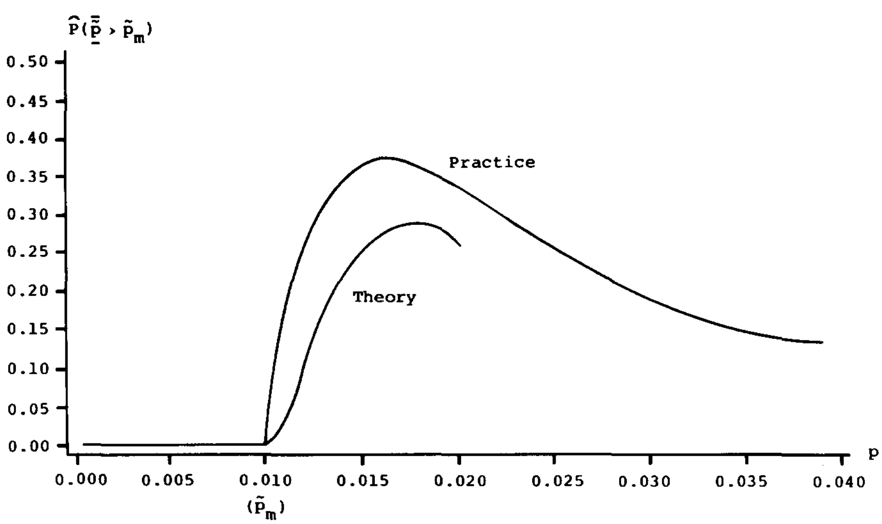

Figure 3. Estimated probability of excessive defectiveness, $P(\bar{\bar{p}} > \tilde{p}_m)$ (NY = 100000; $S = 52$)

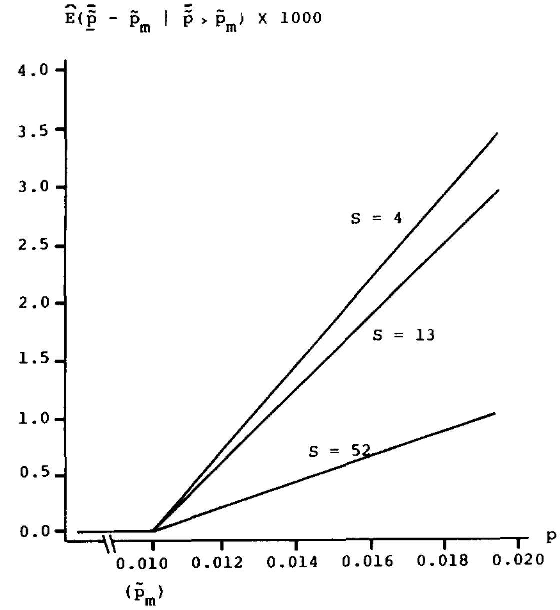

Figure 4. Estimated value $\hat{E}$ of excessive defectiveness, $E(\bar{\bar{p}} - \tilde{p}_m \mid \bar{\bar{p}} > \tilde{p}_m)$ in theoretical approach (NY = 1000000)

is important; the paper suggests a simple estimation scheme based on the AOQL scheme itself (Section 3).

2. Design of Monte Carlo experiment

Table 2 illustrated that the sample size n and the acceptance number k_0 are completely determined by the subpopulation size N_s, the before-sampling fraction p, and the defectiveness limit $\tilde{p}_m$. That subpopulation size N_s depends on the estimated yearly population size NY and on the number of subperiods S. In the simulation we study three values for S, namely 4, 13, and 52 which correspond to quarters, 'months', and weeks; these periods are traditional in accounting practice. The magnitude of NY in the simulation is based on our experience with auditing applications: NY is 10000 or 100000 or 1000000. We assume that $\underline{N}_s$ is uniformly and independently distributed with expected value NY/S; the range is such that the coefficient of variation is roughly 6%, which is an arbitrarily selected value. Note that the actual yearly amount $\Sigma_1^S \underline{N}_s$ deviates from the estimate NY, with probability one.

We further select the following six values for the defectiveness limit $\tilde{p}_m$: 0.1%, 0.5%, 1%, 2%, 5%, 10%. Selection of the before-sampling fraction p in the simulation should relate to the defectiveness limit $\tilde{p}_m$, which can be seen as follows. If p were very high, then the AOQL scheme would be futile: sampling would usually be followed by inspection of the whole subpopulation. Therefore we restrict the simulation to $p \leq 6\tilde{p}_m$. There are no tables available for $p > 2\tilde{p}_m$. This, however, is no problem if only the left most columns of the tables are used (see Section 1). Obviously not all *sub*populations have the

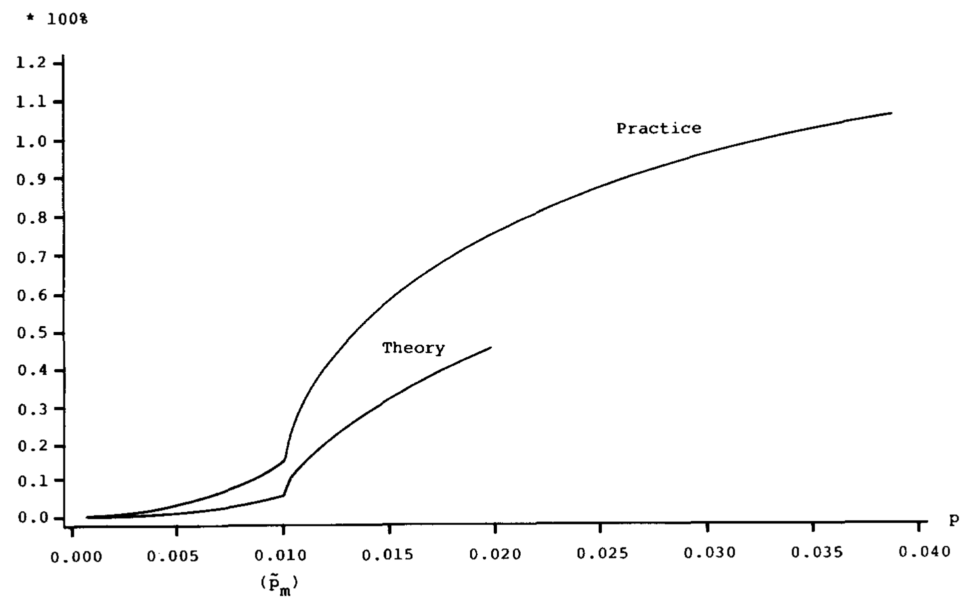

Figure 5. Estimated fraction of subpopulations, fully inspected (NY = 1000000; S = 52)

same p, even if all subpopulations have the same expected value $E(p)$. Therefore we sample p. Figure 1 demonstrated that the performance $E(\bar{\bar{p}})$ improves as p deviates from the least favorable value p^*. In preliminary simulation experiments we sampled p from a distribution with a high variance, and indeed $\bar{\bar{p}}$ decreased (not further reported in this paper). Therefore we concentrate the simulation on *worst cases*: p has a range of only 0.2 $\tilde{p}_m$ (several distributions of p are discussed in Case and Keats, 1982). We further assume that p is uniformly distributed over that range. We do change the expected value $E[p]$: p varies between 0 and $6\tilde{p}_m$ as we explained above. So we sample p from the uniform distribution between 0 and $0.2\tilde{p}_m$, between $0.2\tilde{p}_m$ and $0.4\tilde{p}_m, \ldots$, between $5.8\tilde{p}_m$ and $6\tilde{p}_m$. Figures 2 through 5 do not extend to $p = 6\tilde{p}_m$ because the pattern is clear from figures for smaller values of p.

In total we simulate 1620 factor combinations while using only the left most columns of the tables; this we call the 'practitioner's approach', which is abbreviated to 'Practice' in Figures 2, 3, and 5. We simulate 540 combinations with the optimal $[n, k_0]$ combinations: 'theoretical approach', abbreviated to 'Theory' in these figures.

There is an important technical issue in the simulation: how often (how many years) should each factor combination be simulated in order to obtain *reliable estimates* of performance criteria such as $P[\bar{\bar{p}} > \tilde{p}_m]$? By definition, one replication (one simulated year) yields a binomial variable (say) $\underset{\sim}{x}$ with $q = P(\underset{\sim}{x} = 0) = P[\bar{\bar{p}} > \tilde{p}_m]$. When the normal approximation to the binomial distribution is used, it is straightforward to derive J, the number of simulated years needed to estimate q with either a relative precision of 10% or an absolute precision of 0.001; see Kleijnen (1987, pp. 46–51). We stop as soon as one of these requirements is satisfied. This approximation shows that we need *at most* 16221 replications to satisfy either the relative precision or the absolute precision requirement, with a one-sided probability of 10%; this maximum occurs when $q = 0.01$. Actually we do not know q. So we substitute the 'current' estimate of q after at least 100 replications; that is, we substitute the estimate $\hat{q}_j$ available after j replications with $j = 100, 101, \ldots, J$. The average number of replications turns out to be roughly 1000. We examine not only the performance criterion $q = P[\bar{\bar{p}} > \tilde{p}_m]$, but several more criteria. Yet, since the main criterion is q, we concentrate on q to select the number of replications. The next section will show that the simulation results show patterns not obscured by too much noise.

It takes 40 hours of computer time on a VAX-780 minicomputer to simulate 1620 plus 540 factor combinations, each combination replicated roughly 1000 times. We would have needed even more computer time, had we not introduced the following approximation. The number of defectives $\underset{\sim}{k}$ has a hypergeometric distribution (see Section 1: sampling without replacement). The binomial distribution (sampling with replacement) gives a good approximation provided $n \ll N_s$, which is often the case (but not always: if NY is small, then it may happen that $n > N_s$); see Table 2. In turn, the Poisson distribution provides a good approximation to the binomial distribution if p is small; see Schilling (1982, p. 64). We use the latter approximation, simulating the Poisson distribution through the subroutine in Naylor et al. (1966, p. 114). This Poisson program runs 20 times faster than the hypergeometric program does on our computer.

For completeness sake we mention that we use the multiplicative congruential pseudorandom number generator with multiplier 13^{13} and modulus 2^{59}. This generator was developed and tested by NAG (Numerical Algorithms Group) in the United Kingdom.

3. Monte Carlo results

The Monte Carlo experiment yields an enormous amount of data. We analyze these data through regression analysis (using SAS), in order to smooth the observations and to obtain succinct representations. Preliminary plots looked like gamma functions. Therefore we fit such a nonlinear regression model for the yearly outgoing fraction $\bar{\bar{p}}$ versus the original fraction p, which yields Figure 2 (where 'Practice' refers to using only the left most columns of the tables, and 'Theory' refers to the optimal (n, k_0) combinations; see Section 2). The regression model has an R^2 adjusted for the number of explanatory variables that is higher than 0.95. Figure 2 looks like the theoretical Figure 1: there is a least favorable value for p and $\bar{\bar{p}}$ remains below $\tilde{p}_m$. This result is not surprising, but it *verifies* the correctness of our simulation program!

If the original fraction p satisfies $p \leq \tilde{p}_m$, then obviously $q = P[\bar{\bar{p}} > \tilde{p}_m] = 0$. If, however, $p > \tilde{p}_m$, then we again fit a function like the gamma function, which yields Figure 3. Again R^2 is high: $R^2 = 0.99$ for the theoretical approach, and 0.74 for the practitioner's approach. Figure 3 shows that there is a *sizable probability* of violating the limit on the defectiveness, if the 'practitioner's approach' is followed. The worst case is an estimated probability of 0.618 for $p = 0.017$ (this is one of the observations to which the curve is fitted). We repeat, however, that the simulation concerns *worst cases* (since the fraction p of the subpopulation is sampled from a uniform distribution with a *small* range; see Section 2).

If $\bar{\bar{p}} > \tilde{p}_m$, then *how bad* is the excessive defectiveness $E[\bar{\bar{p}} - \tilde{p}_m \mid \bar{\bar{p}} > \tilde{p}_m]$? Figure 4 shows that smaller *subperiods* (higher S) give extra protection. Our explanation follows from Table 2: if the subpopulation size N is halved (say, from 2000 to 1000 units), then the sample size n decreases only slightly (from 36 to 35); so if the number of subpopulations S increases, then N_s decreases, but the total sample size over a *whole* year increases drastically.

Next, we consider the costs of the sampling plans. Specification of cost functions is rather arbitrary, so we use the fraction of the subpopulations that is rejected and fully inspected. (For specific cost functions we refer to Ercan et al., 1974, Hald, 1981, and Schneider et al., 1988.) The AOQL scheme implies that all N_s units (of a subperiod) are inspected if $k > k_0$. Figure 5 shows that the fraction of fully inspected subpopulations increases drastically if $p > \tilde{p}_m$. Obviously the practitioner's approach is more expensive. The curves are hardly affected by S, the number of subperiods (not displayed).

Note that the simulation shows that it is important to have a good *estimate of* p, the before-sampling fraction of defectives. We might use the estimator $\hat{p} = \underline{k}/n$ if $k \leq k_0$, and $\hat{p} = \underline{K}/\underline{N}_s$ if $k > k_0$ where $\underline{K}$ denotes the number of defectives in the subpopulation (of size $\underline{N}_s$). As time t goes on, we obtain a series of estimators $\underline{\hat{p}}_t$, which can be combined; for example, we may weigh $\underline{\hat{p}}_t$ with the sample size n_t if $k \leq k_0$ and the subpopulation size N_t if $k > k_0$. If $\underline{\hat{p}}_t$ shows serial correlation or non-stationary behavior, we may apply time series techniques. A different approach uses prior distributions; it is discussed by Hald (1981, pp. 15–21, 125–138, 335, 424–425). Since we did not investigate our procedure for estimating p, we do not know if our heuristic is better than Hald's approach is.

4. Conclusions

AOQL sampling plans are indeed used in practice, including auditing. It might be assumed that if the *expected* yearly fraction of defectives after inspection and correction $E(\bar{\bar{p}})$ meets the limit on defectiveness $\tilde{p}_m$, then the *probability* of exceeding that limit $\tilde{p}_m$ is negligible:

$$P\left[\underline{\bar{\bar{p}}} > \tilde{p}_m\right] \approx 0.$$

However, simulation data analyzed by regression models yielded Figure 3, which shows that this probability is sizable if the before-sampling fraction p is higher than the limit $\tilde{p}_m$ but not extremely high (if $p < \tilde{p}_m$, then obviously $\bar{\bar{p}}$ cannot exceed $\tilde{p}_m$; if $p \gg \tilde{p}_m$, then sampling is usually followed by inspection of the whole subpopulation). If p varies much over subperiods, then $P[\bar{\bar{p}} > \tilde{p}_m]$ decreases (we simulated worst case situations: small ranges of p). Figure 4 shows that increasing the number of periods S decreases the magnitude of the expected constraint violation. Underestimating p is not wise: it does not give extra protection (in Figure 3 the 'Practice' curve lies above the 'Theory' curve); yet more inspection work is done (Figure 5). So in practice one should obtain more information about p. One might get estimates of p from the sampling procedure itself: if $k \leq k_0$, then $\hat{p} = \underline{k}/n$; else $\hat{p} = \underline{K}/\underline{N}_s$. To reduce and control $\bar{\bar{p}}$ itself means that the inspection costs decrease (Figure 5); the expected value of the excessive defectiveness also decreases (Figure 4). If p approaches the least favorable value p^* from above, then the probability of excessive defectiveness increases (Figure 3) and the average quality deteriorates (Figure 2). The drive towards zero defects ($p = 0$) gives best results.

Acknowledgments

We thank three anonymous referees for comments that lead to a better presentation.

References

Case, K.E., and Keats, J.B. (1982), "On the selection of a prior distribution in Bayesian acceptance sampling", *Journal of Quality Technology* 14, 10–18.

Cross, R. (1984), "Parts per million AOQL sampling plans", *Quality Progress* 16, 28–34.

Dodge, H.F., and Romig, H.G. (1959), *Sampling Inspection Tables*, second edition, Wiley, New York.

Ercan, S.S., Hassan, M.Z., and Taulananda, A. (1974), "Cost minimizing single sampling plans with AIQL and AOQL constraints", *Management Science* 20, 1112–1121.

Hald, A. (1981), *Statistical Theory of Sampling Inspection by Attributes*, Academic Press, London.

Kleijnen, J.P.C. (1987), *Statistical Tools for Simulation Practitioners*, Marcel Dekker, New York.

Kriens, J., and Veenstra, R.H. (1985), "Statistical sampling in internal control by using the AOQL-system", *The Statistician* 34, 383–390.

Naylor, T.H., Balintfy, J.L., Burdick, D.S., and Chu, K. (1966), *Computer Simulation Techniques*, Wiley, New York.

Schilling, E.G. (1982), *Acceptance Sampling in Quality Control*, Marcel Dekker, New York.

Schneider, H., O'Cinneide, C., and Tang, K. (1988), "Optimal control of a production process subject to AOQL constraint", *Naval Logistics Quarterly* 35, 383–395.

Van Batenburg, P.C., Kriens, J., and Veenstra, R.H. (1988), "Average Outgoing Quality Limit - a revised and improved version", in: J.G. de Gooijer, M.J.T.J. van Nieuwburg and J.A.M. Wesseling (eds.), *Economische Statistiek: Ontwikkelingen in Kwantitatief Onderzoek*, Boom, Meppel, (In Dutch).

Wurnik, F.M. (1984), "Incoming inspection and delivery quality of integrated circuits", *Microelectronic Reliability* 24, 925–933.