PROBABILITY
AND
MATHEMATICAL STATISTICS

Vol. 6, Fasc. 2 (1985), p. 131-137

# ESTIMATION OF A NUMBER OF ERRORS IN CASE OF REPETITIVE QUALITY CONTROL

BY

J. MIELNICZUK (WARSZAWA)

*Abstract*. The estimation of a number of defects of a specified part of a homogeneous product is considered. A natural estimator, although well justified by a heuristical reasoning, is proved to be asymptotically biased. This leads to the proposal of a modified asymptotically unbiased estimator. The asymptotic variances of both estimator are derived and compared with the results of a Monte-Carlo study.

**1. Introduction.** We consider the following scheme of repetitive quality control. Two controllers are seeking independently for defects of a homogeneous product, e.g. a bale of cloth. They check $t$ units of length. Assume that the defects are randomly distributed on the material so that the number of defects in $t$ units, denoted by $n$, has the Poisson distribution $P(\gamma t)$; $\gamma$ is a positive constant describing defectiveness of the whole product. Defectiveness $\gamma$ and probabilities $p_1$ and $p_2$ of finding a single defect by the respective controller are unknown. We assume that $0 < p_1, p_2 < 1$. We observe $n_1, n_2, m$ where $n_i$ $(i = 1, 2)$ is the number of defects found by the $i$-th controller in $t$ units and $m$ is the number of defects found by both controllers simultaneously.

We are interested in estimating $n$ when $n$ is fixed but large or in estimating $\gamma$ when $t$ is fixed but large. To avoid misunderstandings we use the notation $E_n \hat{n}$ and $\mathrm{Var}_n \hat{n}$ in the first case and $E_t \hat{\gamma}$ and $\mathrm{Var}_t \hat{\gamma}$ in the second case.

It is obvious that any reasonable estimator $\hat{n}(n_1, n_2, m)$ of $n$ should fulfil the condition

$$\hat{n} > n_1 + n_2 - m. \tag{1}$$

Polya [2] introduced the following heuristically justified estimator:

$$\hat{n} = \begin{cases} n_1 n_2/m, & m > 0, \\ n_1 + n_2, & m = 0. \end{cases} \tag{2}$$

It is based on the following idea: since $n_1 \sim Bin(n, p_1)$, $n_2 \sim Bin(n, p_2)$, $m \sim Bin(n, p_1 p_2)$, and the respective expected values are $np_1$, $np_2$ and $np_1 p_2$, the ratio $n_1 n_2/m$ should be close to $np_1 \cdot np_2/np_1 p_2$, the later being equal to $n$. Condition (1) is satisfied since $0 \leqslant m \leqslant \min(n_1, n_2)$ and whence $n_1 n_2 \geqslant (n_1 + n_2) m - m^2$.

In this note we show that despite of the heuristic justification, $n$ is asymptotically biased with bias equal to $1/p_1 p_2 + 1 - 1/p_1 - 1/p_2$. Thus we introduce the modified estimator $n$:

$$\tilde{n} = \begin{cases} \hat{n} - \dfrac{n_1 n_2}{m^2} + \dfrac{n_1}{m} + \dfrac{n_2}{m} - 1, & m > 0, \\ n_1 + n_2, & m = 0, \end{cases} \tag{3}$$

which proves to be asymptotically unbiased, We also calculate the asymptotic variance of $\tilde{n}$. Moreover, we propose the natural estimator of $\gamma$, $\tilde{\gamma} = \tilde{n}/t$, and study its asymptotic properties.

## 2. Main result.

THEOREM 1. *Let*

$$p = \frac{1}{p_1 p_2} + 1 - \frac{1}{p_1} - \frac{1}{p_2}.$$

*Then*

$$\lim_{n \to \infty} (\mathrm{E}_n \hat{n} - n) = p, \tag{4}$$

$$\lim_{n \to \infty} (\mathrm{E}_n \tilde{n} - n) = 0, \tag{5}$$

$$\lim_{n \to \infty} (\mathrm{Var}_n \hat{n} - np) = a_1(p_1, p_2), \tag{6}$$

$$\lim_{n \to \infty} (\mathrm{Var}_n \tilde{n} - np) = a_2(p_1, p_2), \tag{7}$$

*where*

$$a_1(p_1, p_2) = 2p^2 + \frac{5}{p_1 p_2} + \frac{1}{p_1^2} + \frac{1}{p_2^2} - \frac{1}{p_1^2 p_2^2} - 1,$$

$$a_2(p_1, p_2) = p^2 + \frac{1}{p_1 p_2} p.$$

Note that the terms of order $n$ of variances in (6) and (7) are equal. It is easy to see that the asymptotic bias $p$ as well as $a_i$ $(i = 1, 2)$ and $a_1 - a_2$ are positive and unbounded from above on an open quadrat $(0, 1) \times (0, 1)$ and tend to 0 when $p_1$ and $p_2$ tend to 1.

In order to prove Theorem 1, let us start with the following two lemmas. To simplify the notation we omit subscript $n$ in $E_n$.

LEMMA 1. *Let $X \sim Bin(n, p_1)$, where $0 < p_1 < 1$. Then for every natural $n$*

$$\left| E\left(\frac{1}{X} \mid X > 0\right) - \frac{1}{p_1(n+1)} - \frac{1}{p_1^2(n+1)(n+2)} \right| \leqslant k/n^3,$$

*where $k$ is a positive constant not depending on $n$.*

Proof. First observe that

$$\frac{1}{x} = \frac{1}{x+1} + \frac{1}{(x+1)(x+2)} + \frac{2}{x(x+1)(x+2)}$$

and

$$\frac{2}{x(x+1)(x+2)} \leqslant \frac{8}{(x+1)(x+2)(x+3)} \quad \text{for } x \geqslant 1.$$

By easy computations it can be shown that (with $E^*(f(X))$ denoting $E(f(X) \mid X > 0)$ for any $f$)

$$E^*\left(\frac{1}{X+1}\right) = \frac{a}{p_1(n+1)}(1 - q_1^{n+1} - (n+1)p_1 q_1^n),$$

$$E^*\left(\frac{1}{(X+1)(X+2)}\right)$$

$$= \frac{a}{p_1^2(n+1)(n+2)}\left(1 - q_1^{n+2} - (n+2)p_1 q_1^{n+1} - \frac{1}{2}(n+1)(n+2)p_1^n q_1^2\right),$$

$$E^*\left(\frac{1}{(X+1)(X+2)(X+3)}\right) \leqslant \frac{a}{p_1^3(n+1)(n+2)(n+3)},$$

where $a = (1-q_1^n)^{-1}$ and $q_1 = 1 - p_1$. Using the fact that $q_1^n$ and $(1-a)$ are both of an order less than $n^{-3}$, we have

$$E^*\left(\frac{1}{X+1}\right) + E^*\left(\frac{1}{(X+1)(X+2)}\right) + E^*\left(\frac{8}{(X+1)(X+2)(X+3)}\right)$$

$$= \frac{1}{p_1(n+1)} + \frac{1}{p_1^2(n+1)(n+2)} + o(n^{-3}).$$

Consequently, the proof is completed by the triangle inequality.

Note that Lemma 1 is a generalization of Lemma 4.2 in [1].

LEMMA 2. *Let $n_2' = n_2 - m$. For every natural $i$ the random variables $(n_2' \mid m > 0 \wedge n_1 = i)$ and $(m \mid m > 0 \wedge n_1 = i)$ are independent and have distributions $Bin(n-i, p_2)$ and $Bin(i, p_2)$, respectively.*

The proof is immediate.

Proof of (4). We have

$$\mathrm{E}(\hat{n}) = P(m=0)\cdot \mathrm{E}(\hat{n}|m=0)+\sum_{i=1}^{n} P(n_1 = i \wedge m>0)\cdot \mathrm{E}(\hat{n}|n_1 = i \wedge m>0)$$

$$= o(1)+\sum_{i=1}^{n} \mathrm{P}(m>0|n_1 = i)\cdot \mathrm{P}(n_1 = i)\cdot \mathrm{E}(\hat{n}|n_1 = i \wedge m>0).$$

The last equality holds, since

$$0 \leqslant \mathrm{P}(m=0)\cdot \mathrm{E}(\hat{n}|m=0) \leqslant n\cdot(1-p_1 p_2)^n.$$

Note that

(8) $$\sum_{i=1}^{n} (1-p_2)^i \,\mathrm{P}(n_1 = i)\, E\left(\frac{n_1 n_2}{m}|n_1 = i \wedge m>0\right) = o(1).$$

Since

$$(1-p_2)^i = o(i^{-3}), \qquad \mathrm{E}\left(\frac{n_1 n_2}{m}|n_1 = i \wedge m>0\right) \leqslant ni,$$

the sum in (8) is less than (see [1])

$$\frac{n}{a}\mathrm{E}^*\left(\frac{1}{X^2}\right) = o(1).$$

Therefore

$$\mathrm{E}(\hat{n}) = o(1)+\sum_{i=1}^{n} [1-(1-p_2)^i]\,\mathrm{P}(n_1 = i)\cdot \mathrm{E}\left(\frac{n_1 n_2}{m}|n_1 = i \wedge m>0\right)$$

$$= o(1)+\sum_{i=1}^{n} \mathrm{P}(n_1 = i)\,\mathrm{E}\left(\frac{n_1 n_2}{m}|n_1 = i \wedge m>0\right).$$

By Lemma 2

$$\mathrm{E}\hat{n} = o(1)+\sum_{i=1}^{n} \binom{n}{i} p_1^i (1-p_1)^{n-i}\cdot i\cdot\left(1+(n-i)\cdot p_2\right)\cdot \mathrm{E}^*\left(\frac{1}{m}|n_1 = i\right)$$

and, by Lemma 1, neglecting again the terms of order $i^{-3}$, we have

$$\mathrm{E}\hat{n} = o(1)+\sum_{i=1}^{n} \binom{n}{i} p_1^i (1-p_1)^{n-i} i\cdot\left(1+p_2(n-i)\right)\times$$

$$\times\left\{\frac{1}{p_2 i}-\frac{1}{p_2 i(i+1)}+\frac{1}{p_2^2 (i+1)(i+2)}\right\}.$$

Summing the first term in curly brackets we get

(9) $$n-nq_1^n = n+o(1).$$

For the second and third terms we get, respectively,

$$(1-q_1^n)\left(-(n+1)\mathrm{E}^*\left(\frac{1}{X+1}\right)+1\right) = -\frac{1}{p_1}+1+o(1) \tag{10}$$

and

$$\sum_{i=1}^{n} \mathrm{P}(n_1 = i)\cdot\frac{(n-1)(i+2)-2(n-i)}{p_2(i+1)(i+2)} \tag{11}$$

$$= (1-q_1^n)\cdot\left(\frac{(n+1)}{p_2}\mathrm{E}^*\left(\frac{1}{X+1}\right)-\frac{1}{p_2}+o(1)\right) = \frac{1}{p_1 p_2}-\frac{1}{p_2}+o(1).$$

Equation (4) follows from (9)-(11).

Proof of (5). It is enough to show that

$$\lim_{n\to\infty} \mathrm{E}\frac{n_1}{m} = \frac{1}{p_1}, \quad \lim_{n\to\infty} \mathrm{E}\frac{n_2}{m} = \frac{1}{p_2}, \tag{12}$$

$$\lim_{n\to\infty} \mathrm{E}\frac{n_1 n_2}{m^2} = \frac{1}{p_1 p_2}. \tag{13}$$

Equations (12) are simple consequences of Lemma 1. As for (13), we prove the inequality

$$\left|\mathrm{E}\left(\frac{1}{X^2}|X>0\right)-\frac{1}{p_1^2(n+1)(n+2)}\right| \leqslant k/n^3$$

in the same way as Lemma 1. Thus (cf. (11))

$$\mathrm{E}\frac{n_1 n_2}{m^2} = \mathrm{E}\frac{n_1}{m}+\sum_{i=1}^{n} \mathrm{P}(n_1 = i)\frac{i(n-i)}{p_2(i+1)(i+2)}+o(1)$$

$$= \frac{1}{p_2}+\frac{1}{p_2}\left(\frac{1}{p_1}-1\right)+o(1) = \frac{1}{p_1 p_2}+o(1).$$

The proofs of (6) and (7) are based on Lemma 2 and the expansions of suitable order of $\mathrm{E}(1/X^i)$, where $i = 1, 2, 3$ and $X \sim Bin(n, p)$. The approach is similar to that used in proofs of (4) and (5) and therefore we omit the details.

THEOREM 2. $\mathrm{E}_t\tilde{\gamma} = \gamma+o(1/t)$, $\mathrm{Var}_t\tilde{\gamma} = \gamma/t+a_2/t^2+o(1/t^2)$.

The proof of Theorem 2 follows from Theorem 1 and the formulas

$$t\mathrm{E}_t\tilde{\gamma} = \sum_{k=0}^{\infty} p_k \mathrm{E}_k \tilde{n},$$

$$t^2\,\mathrm{Var}_t\tilde{\gamma} = \sum_{k=0}^{\infty} p_k \mathrm{Var}_k\tilde{n}+\sum_{k=0}^{\infty} p_k(\mathrm{E}_k\tilde{n}-\mathrm{E}_t\tilde{n})^2$$

where $p_k = (\gamma t)^k e^{-\gamma t}/k!$.

It is easy to see that the bias of the estimator $\hat{\gamma} = \hat{n}/t$ is equal to $p/t + o(1/t)$, while the asymptotic variance of $\hat{\gamma}$ is similar to that of $\tilde{\gamma}$ with $\gamma/t$ and $a_2$ replaced by $\gamma(1+p)/t$ and $a_1$, respectively. Thus the main term of the asymptotic variance of $\tilde{\gamma}$ is smaller than that of the asymptotic variance of $\hat{\gamma}$.

Since the formulas for expectations and variances are asymptotic, a Monte-Carlo study has been performed for various $n, p_1, p_2$. The approximation of expectation and variances seems satisfactory for $p_1, p_2 \geqslant 0.5$ and $n \geqslant 50$. For such $n, p_1, p_2$

$$\frac{|\Sigma_{AS}(\tilde{n}) - \Sigma_{SM}(\tilde{n})|}{n} \leqslant \frac{1}{50} \quad \frac{|E_{AS}(\tilde{n}) - E_{SM}(\tilde{n})|}{n} \leqslant \frac{1}{100},$$

where subscripts AS and SM stand for "asymptotic" and "simulated", and $\Sigma$ denotes standard deviation. The same inequalities are satisfied when $\tilde{n}$ is replaced by $\hat{n}$. The simulation results for $n = 50$ are given in Table I.

Table I. Asymptotic and simulated means and variances of $\hat{n}$ and $\tilde{n}$ for $n = 50$

| | | $E(\hat{n})$ | $E(\tilde{n})$ | $\Sigma(\hat{n})$ | $\Sigma(\tilde{n})$ |
|---|---|---|---|---|---|
| $p_1 = p_2 = 0.5$ | SM | 51.238 | 49.796 | 8.148 | 6.435 |
| | AS | 51 | 50 | 7.937 | 7.416 |
| $p_1 = p_2 = 0.7$ | SM | 50.201 | 49.982 | 3.098 | 2.818 |
| | AS | 50.184 | 50 | 3.168 | 3.096 |
| $p_1 = p_2 = 0.9$ | SM | 50.019 | 50.005 | 0.786 | 0.746 |
| | AS | 50.012 | 50 | 0.793 | 0.8 |

The approximation is less satisfactory outside this region. For example, for $p_1 = p_2 = 0.3$ and $n = 100$ simulated (asymptotic) mean value and standard deviation of $\tilde{n}$ are 93.004 (100) and 20.480 (25.179), respectively. For $p_1 = p_2 = 0.5$ and $n = 40$ we have $\Sigma_{SM}(\tilde{n}) = 8.814$ and $\Sigma_{AS}(\tilde{n}) = 7.280$. Similar situation can occur when $p_1$ or $p_2$ is less than 0.5; e.g. for $p_1 = 0.2$, $p_2 = 0.7$ and $n = 100$ we have 13.243 for the simulated standard deviation of $n$ and 16.492 for the asymptotic one.

**Acknowledgments.** The author is indebted the referee for his remarks concerning the earlier version of this paper and J. Ćwik for his help in performing the Monte-Carlo experiment.

## BIBLIOGRAPHY

[1] O. Aalen, *Nonparametric inference in connection with multiple decrement models*, Scand. J. Statist. 3 (1976), p. 15-27.

[2] G. Polya, *Probabilities in proofreading*, Am. Math. Monthly 83 (1976), p. 42.

Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland

*Received on 17. 6. 1982;*
*revised version on 20. 11. 1984*