ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XIII, 3 (1973)

R. ZIELIŃSKI (Warszawa)

A NEW CLASS OF ESTIMATORS WITH AN APPLICATION TO STATISTICAL QUALITY CONTROL

The following problem is typical when the quality of liquids, powders and the like materials is to be controlled. A barrel of a liquid is delivered. The liquid is polluted by a substance. The amount Z of the polluting substance in the barrel is to be estimated but only a small quantity of the liquid (s-sample) can be measured. If the liquid is or may be thoroughly mixed, any "chung" will be a good sample. If the polluting substance is not uniformly dispersed in the liquid and the mixing before sampling is not possible, the sample should be taken at random. The random sampling will be satisfactory if the expected value of what is observed in the sample is equal to Z (unbiasedness) and the variance of sample values is small (effectiveness).

Such problems arise when the quality of soil, ores and the like stuffs is investigated, when the amount of a given product in a mixture (e. g. in an alloy) is to be estimated, when water pollution or air pollution is to be measured, etc.

Consider the one-dimensional variant of such problems. The abovementioned problem of the estimation of Z can be considered as one-dimensional if, for example, the liquid stratifies and the concentration of the polluting substance in any horizontal section of the barrel is constant.

The one-dimensional problem can be formalized as follows: evaluate the integral

$$I = \int_0^1 f(t)\,dt$$

using an estimator like

$$\hat{I}(X) = \int_{X-\delta}^{X+\delta} f(t)\,dt, \tag{1}$$

where δ is a given number and X is a random variable.

In the case of a stratified liquid this formulation can be justified by the following discussion. Let the liquid fill a cylindric vessel (Fig. 1). Let the height of the vessel be equal to 1 and the area of its base be equal to S. The value $f(t)$ is interpreted as the concentration of the polluting substance at the point t $(0 \leqslant t \leqslant 1)$ so that $S\int_0^1 f(t)\,dt$ is equal to the total amount of the polluting substance in the vessel. The sample (lined in the figure) forms a cylinder with a base of any arbitrary shape and the area equal to s. The amount of the polluting substance in the sample is equal to $s\int_{X-\delta}^{X+\delta} f(t)\,dt$. The number δ is proportional to the sample size and the random variable X indicates the location of the sample. We shall assume $\delta \in (0, \frac{1}{2})$.

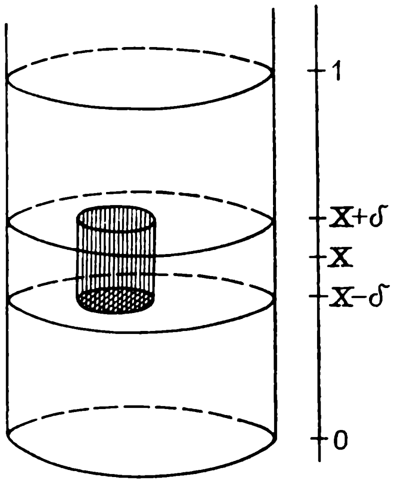

Fig. 1

Usually samples are taken "at random"; then X is uniformly distributed on the interval $(\delta, 1-\delta)$. In this case the estimator (1) will be denoted by $\hat{I}_1$. We shall show that $\hat{I}_1$ is a biased estimator, i. e. $\mathrm{E}\,\hat{I}_1 \neq I$. That is why two others estimators $\hat{I}_2$ and $\hat{I}_3$ which are unbiased will be proposed. The estimator $\hat{I}_2$ is constructed as a modification of $\hat{I}_1$ and its unbiasedness is achieved by change in sample size at the ends of the interval $(0, 1)$. The estimator $\hat{I}_3$ is constructed for X with a discrete distribution. The well-known technique of importance sampling enables us to construct $\hat{I}_2$ and $\hat{I}_3$ which variances are very small.

* * *

Let P be the distribution function of the random variable X. Then the expected value of (1) is given by the obvious formula

$$\mathrm{E}\,\hat{I}(X) = \frac{1}{2\delta}\int_{\delta}^{1-\delta}\int_{x-\delta}^{x+\delta} f(t)\,dt\,dP(x).$$

For the estimator $\hat{I}_1$, we have $dP(x) = dx/(1-2\delta)$. Hence

$$\mathrm{E}\,\hat{I}_1 = \frac{\lambda}{1-\lambda}I + \frac{\int_0^{\lambda}(t-\lambda)f(t)\,dt + \int_{1-\lambda}^{1}(1-\lambda-t)f(t)\,dt}{\lambda(1-\lambda)}, \tag{2}$$

where $\lambda = \min(2\delta, 1-2\delta)$. From (2) we conclude that $\hat{I}_1$ is a biased estimator of I. If $\delta \to 0$ or $\delta \to \frac{1}{2}$, the estimator $\hat{I}_1$ becomes an unbiased estimator. In the former case, $\hat{I}(X)$ tends to a standard estimator $f(X)$; in the latter one, the sample is formed by the whole "population".

Let X have the distribution with a density function $g(x) > 0$ for $0 < x < 1$, and write

$$\hat{I}_2(x) = \begin{cases} \dfrac{1}{2\delta g(x)} \displaystyle\int_0^{x+\delta} f(t)\,dt, & 0 < x < 2\delta, \\ \dfrac{1}{2\delta g(x)} \displaystyle\int_{x-\delta}^{x+\delta} f(t)\,dt, & 2\delta \leqslant x < 1-2\delta, \\ \dfrac{1}{2\delta g(x)} \displaystyle\int_{x-\delta}^{1} f(t)\,dt, & 1-2\delta \leqslant x < 1. \end{cases} \tag{3}$$

Let the random variable X have a discrete distribution

$$\Pr\{X = (2k-1)\delta\} = p_k, \quad k = 1, 2, \ldots, m,$$

where m is a positive integer such that $2\delta m = 1$ and $p_k > 0$ for all $k = 1, 2, \ldots, m$. Write

$$\hat{I}_3(x) = \frac{1}{\Pr\{X = x\}} \int_{x-\delta}^{x+\delta} f(t)\,dt. \tag{4}$$

The estimators $\hat{I}_2$ and $\hat{I}_3$ are unbiased for any distribution $g(x)$ or $p_1, p_2, \ldots, p_k$. The proof of this statement is obvious.

Note that, by a suitable choice of the distribution of random variable X, we can let the variances of the estimators to be very small. Assuming

$$p_k = C \int_{x-\delta}^{x+\delta} f(t)\,dt, \quad k = 1, 2, \ldots, m, \tag{5}$$

where C is a constant (such that $\sum p_k = 1$), we have $D^2\hat{I}_3 = 0$. Such a procedure is of course not available in practical problems because there is no problem of estimation of I when all the integrals in (5) are known. But choosing the distribution of X in a reasonable way, we can reduce considerably the variance of the estimator. For example, if $f(x)$ is a decreasing function, then $g(x)$ should be taken as a decreasing function or $p_1, p_2, \ldots, p_k$ as a decreasing sequence. This technique is well known as importance sampling.

In practical problems, X usually has the uniform distribution; then

$g(x) \equiv 1$ and $p_k \equiv 2\delta$. It is interesting that then $D^2\hat{I}_2 > 0$ even if $f(x) = \mathrm{const}$ (for $f(x) \equiv 1$, we have $D^2\hat{I}_2 = \delta/3$).

INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES

Received on 13. 10. 1972

R. ZIELIŃSKI (Warszawa)

NOWY TYP ESTYMATORÓW I PEWNE ICH ZASTOSOWANIE W STATYSTYCZNEJ KONTROLI JAKOŚCI

STRESZCZENIE

W pracy rozpatruje się estymatory postaci (1) dla oszacowania całki $\int_0^1 f(t)\,dt$. Estymatory tego typu znajdują zastosowanie w statystycznej kontroli jakości tzw. ciał bezkształtnych (płyny, proszki itd.). Estymator (1), gdy X ma rozkład równomierny w przedziale $(\delta, 1-\delta)$, jest obciążony. Podano dwa estymatory nieobciążone (3) i (4); przez odpowiedni wybór rozkładów zmiennej losowej X dla tych estymatorów można osiągnąć to, że ich wariancja będzie bardzo mała.
