Colloquium Biometricum 46
2016, 9-14

A REMARK ON ESTIMATING DEFECTIVENESS IN SAMPLING ACCEPTANCE INSPECTION

Wojciech Zieliński

Department of Econometrics and Statistics
Warsaw University of Life Sciences
Nowoursynowska 159, PL-02-787 Warszawa
e-mail: wojciech_zielinski@ssggw.pl

Summary

The problem of estimating a probability of success in a Binomial model is considered. The classical estimator is compared with the estimator which uses the information about non-homogeneity of the sample. An application to the problem of estimating defectiveness in sampling acceptance inspection is shown.

Keywords and phrases: binomial model, probability of success

Classification AMS 2010: 62F03, 62P10

One of the aspects of the sampling acceptance investigations provided in statistical quality control is estimating the defectiveness of a batch of goods. Each item is observed whether it is defective or not.The defectiveness is measured by the percentage of the defective items ("off-types"). To do so the sample of size n is taken and the number of off-types is counted. To be more formal, let ξ be a number of off-types in n trials. This is a random variable binomially distributed. The statistical model for ξ is

$$(\{0, 1, \ldots, n\}, \{Bin(n, \theta), \theta \in (0, 1)\})$$

and the unbiased estimator with minimal variance of the parameter θ is $\hat{\theta}_c = \frac{\xi}{n}$. The variance of that estimator equals

$$D^2_\theta \hat{\theta}_c = \frac{\theta(1-\theta)}{n} \text{ for all } \theta.$$

Suppose now that the defectiveness depends on a supplier. For simplicity assume that there are two different suppliers, which results in defectiveness θ_1 and θ_2, respectively. We are interested in estimation the overall defectiveness θ. The question is, does the information of those different suppliers improve the estimation of θ?

Let the contribution of the first supplier be $w_1 \cdot 100\%$. Then the overall defectiveness is

$$\theta = w_1\theta_1 + w_2\theta_2,$$

where $w_2 = 1 - w_1$. Let n_1 and n_2 denote the sample sizes from the first and the second supplier, respectively. The whole sample size equals $n = n_1 + n_2$.

Now we have two random variables

$$\xi_1 \sim Bin(n_1, \theta_1), \quad \xi_2 \sim Bin(n_2, \theta_2),$$

where $\theta = w_1\theta_1 + w_2\theta_2$ and $w_1 + w_2 = 1$.

The values θ_1 and θ_2 as well as θ are unknown. It is easy to note, that for a given θ parameter θ_1 may take on the values from the interval

$$\left(\max\left\{0, \frac{\theta - w_2}{w_1}\right\}, \ \min\left\{1, \frac{\theta}{w_1}\right\}\right).$$

Denote by a_θ the left end of the above interval and by b_θ its right end, i.e.

$$a_\theta = \max\left\{0, \frac{\theta - w_2}{w_1}\right\} \quad \text{and} \quad b_\theta = \min\left\{1, \frac{\theta}{w_1}\right\}.$$

Consider a random variable

$$\hat{\theta}_w = w_1\frac{\xi_1}{n_1} + w_2\frac{\xi_2}{n_2}.$$

It is an unbiased estimator of θ:

$$\begin{aligned} E_\theta\hat{\theta}_w &= E_\theta\left(w_1\frac{\xi_1}{n_1} + w_2\frac{\xi_2}{n_2}\right) \\ &= \frac{1}{b_\theta - a_\theta}\int_{a_\theta}^{b_\theta}\left(\frac{w_1}{n_1}E_{\theta_1}\xi_1 + \frac{w_2}{n_2}E_{\frac{\theta - w_1\theta_1}{w_2}}\xi_2\right)d\theta_1 \\ &= \frac{1}{b_\theta - a_\theta}\int_{a_\theta}^{b_\theta}\left(\frac{w_1}{n_1}n_1\theta_1 + \frac{w_2}{n_2}n_2\left(\frac{\theta - w_1\theta_1}{w_2}\right)\right)d\theta_1 \\ &= \theta. \end{aligned}$$

The variance of that estimator equals:

$$
\begin{aligned}
D^2_\theta\hat{\theta}_w &= D^2_\theta\left(w_1\frac{\xi_1}{n_1}+w_2\frac{\xi_2}{n_2}\right)\\
&=\frac{1}{b_\theta-a_\theta}\int_{a_\theta}^{b_\theta}\left(\frac{w_1^2}{n_1^2}D^2_{\theta_1}\xi_1+\frac{w_2^2}{n_2^2}D^2_{\frac{\theta-w_1\theta_1}{w_2}}\xi_2\right)d\theta_1\\
&=\frac{1}{b_\theta-a_\theta}\int_{a_\theta}^{b_\theta}\left(\frac{w_1^2\theta_1(1-\theta_1)}{n_1}+\frac{w_2^2\frac{\theta-w_1\theta_1}{w_2}\left(1-\frac{\theta-w_1\theta_1}{w_2}\right)}{n_2}\right)d\theta_1\\
&=\frac{1}{b_\theta-a_\theta}\int_{a_\theta}^{b_\theta}\left(\frac{w_1^2\theta_1(1-\theta_1)}{n_1}+\frac{(\theta-w_1\theta_1)\,(1-\theta-w_1(1-\theta_1))}{n-n_1}\right)d\theta_1.
\end{aligned}
$$

It may be written as

$$
\frac{2n_1w_1(b_\theta^3-a_\theta^3)+3w_1(n_1(1-2\theta)-nw_1)(b_\theta^2-a_\theta^2)+6n_1\theta(\theta+w_1-1)(b_\theta-a_\theta)}{6n_1(n-n_1)(b_\theta-a_\theta)}
$$

or (for $w_1\le 0.5$)

$$
D^2_\theta\hat{\theta}_w=\begin{cases}\frac{\theta(3n_1+3nw_1-6n_1w_1-2n\theta)}{6n_1(n-n_1)}, & \text{for } 0<\theta<w_1,\\ \frac{nw_1^2-3n_1w_1+6\theta(1-\theta)}{6n_1(n-n_1)}, & \text{for } w_1\le\theta\le 1-w_1,\\ \frac{(1-\theta)(3n_1+3nw_1-6n_1w_1-2n(1-\theta))}{6n_1(n-n_1)}, & \text{for } 1-w_1<\theta<1.\end{cases}
$$

The question is whether the variance $D^2_\theta\hat{\theta}_w$ is smaller than the variance $D^2_\theta\hat{\theta}_c$. It may be seen that for given w_1 the variance $D^2_\theta\hat{\theta}_w$ depends on n_1 and may be smaller as well as greater than $D^2_\theta\hat{\theta}_c$. We would like to find optimal n_1, i.e. the size of the first sample which gives the minimal variance $D^2_\theta\hat{\theta}_w$ (the overall sample size n is treated as given). With no loss of generality it may be assumed that $w_1\le 0.5$.

It is easy to note, that

1. $D^2_\theta\hat{\theta}_w=D^2_{1-\theta}\hat{\theta}_w$,

2. $D^2_0\hat{\theta}_w=0$,

3. $\max_{\theta\in\langle 0,1\rangle}D^2_\theta\hat{\theta}_w=D^2_{0.5}\hat{\theta}_w$ and $\max_{\theta\in\langle 0,1\rangle}D^2_\theta\hat{\theta}_c=D^2_{0.5}\hat{\theta}_c$.

Hence, it is enough to find the optimal n_1 for $\theta=0.5$, i.e. n_1 such that $D^2_{0.5}\hat{\theta}_w$ attains its minimum.

For $\theta = 0.5$ we have $a(0.5) = 0$ and $b(0.5) = 1$. After some calculations we obtain:

$$D^2_{0.5}\hat{\theta}_w = \int_0^1 \left(\frac{w_1^2\theta_1(1-\theta_1)}{n_1} + \frac{(0.5 - w_1\theta_1)\,(0.5 - w_1(1-\theta_1))}{n-n_1} \right) d\theta_1$$
$$= \frac{0.25n_1 - 0.5n_1w_1 + nw_1^2/6}{n_1(n-n_1)}.$$

The value of n_1^* minimizing the variance is

$$n_1^* = pn, \text{ where } p = \frac{w_1}{w_1 + \sqrt{1.5 - 3w_1 + w_1^2}}.$$

Since n_1^* may not be an integer, so the optimal size n_1^{opt} of the first sample is n_1^* rounded to the nearest integer. Examples of numerical solutions are given in Table 1 (for $n = 100$). For comparison, in the last column of Table 1 the variance $D^2_{0.5}\hat{\theta}_c$ is given.

Table 1. Minimal variances $D^2_{0.5}\hat{\theta}_w$ for $n = 100$

w_1	p	n_1^{opt}	$D^2_{0.5}\hat{\theta}_w$	$D^2_{0.5}\hat{\theta}_c$
0.05	0.0412	4	0.0024523	0.0025
0.10	0.0833	8	0.0024004	0.0025
0.15	0.1265	13	0.0023431	0.0025
0.20	0.1710	17	0.0022797	0.0025
0.25	0.2171	22	0.0022096	0.0025
0.30	0.2653	27	0.0021309	0.0025
0.35	0.3163	32	0.0020412	0.0025
0.40	0.3710	37	0.0019377	0.0025
0.45	0.4312	43	0.0018156	0.0025
0.50	0.5000	50	0.0016667	0.0025

Note that the minimal variances $D^2_{0.5}\hat{\theta}_w$ are smaller than $D^2_{0.5}\hat{\theta}_c$. It may be easily checked that for $n_1 = n_1^{opt}$ the variance $D^2_\theta\hat{\theta}_w$ is smaller than $D^2_\theta\hat{\theta}_c$ for all $\theta \in (0,1)$. In Figure 1 the variances $D^2_\theta\hat{\theta}_w$ and $D^2_\theta\hat{\theta}_c$ are plotted as the functions of probability θ.

In Table 2, the relative error of the estimation is shown. Precisely, one of the indicators of the goodness of the estimator is its relative error. The probability of not making a relative mistake less than given $\varepsilon > 0$

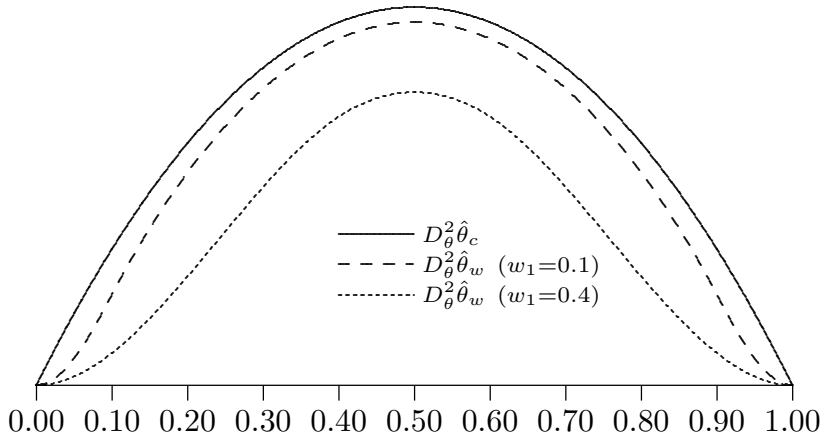

Fig. 1. Variances of estimators for $n = 100$

Table 2. $n = 100$, $n_1 = 37$, $\varepsilon = 0.1$

θ	$P_\theta\left\{\frac{\lvert\hat{\theta}_w-\theta\rvert}{\theta} < \varepsilon\right\}$	$P_\theta\left\{\frac{\lvert\hat{\theta}_c-\theta\rvert}{\theta} < \varepsilon\right\}$
0.01	0.74041	0.73576
0.02	0.54831	0.54407
0.03	0.45685	0.45263
0.04	0.40104	0.39672
0.05	0.36264	0.35816
0.10	0.40293	0.38216
0.15	0.44677	0.42516
0.20	0.49237	0.46774
0.25	0.54143	0.51095
0.30	0.60364	0.55486
0.35	0.66597	0.59685
0.40	0.72788	0.64163
0.45	0.75671	0.68302
0.50	0.79292	0.76770

by considered estimators are calculated. Numerical results are given for $w_1 = 0.4$ and $n_1 = 37$ (which is optimal for $w_1 = 0.4$, see Table 1) and $\varepsilon = 0.1$ (i.e. the relative error does not exceed 10%). It is seen that the probability of "correct" estimation is higher for $\hat{\theta}_w$ than for $\hat{\theta}_c$.

Probabilities given in Table 2 may be of course calculated for other w_1,

n and ε. The distribution of the estimator $\hat{\theta}_w$ is given by the formula:

$$P_\theta\left\{\hat{\theta}_w = u\right\} = \sum_{x_1=0}^{n_1} P_\theta\left\{\xi_1 = x_1, \xi_2 = \frac{n_2}{w_2}\left(u - w_1\frac{x_1}{n_1}\right)\right\}.$$

The probability $P_\theta\{\xi_1 = x_1, \xi_2 = x_2\}$ for $x_1 = 0, 1, \ldots, n_1$ and $x_2 = 0, 1, \ldots, n_2$ equals

$$\frac{1}{b_\theta - a_\theta}\int_{a_\theta}^{b_\theta} f(x_1; n_1, \theta_1)\, f\left(x_2; n_2, \frac{\theta - w_1\theta_1}{w_2}\right) d\theta_1,$$

and equals 0 elsewhere. Here

$$f(x; n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n.$$

We consider the relative error of the estimator. It seems quite natural to consider the absolute error, i.e. $|\hat{\theta}_c - \theta|$ or $|\hat{\theta}_w - \theta|$. This kind of error is not a good measure of goodness in the problem of estimation of the probability of success, because the binomial distribution for small (and large) values of θ is highly skew (its coefficient of skewness equals $(1-2\theta)/\sqrt{n\theta(1-\theta)}$).

The advantages of the estimator $\hat{\theta}_w$ were shown. Hence in practice, it is recommended to use the information of different defectiveness. It will be interesting to generalize the results above to more than two "subpopulations". The work on the subject is in progress.

Bibliographical note. All necessary information on binomial distribution and the estimation of probability of success may be found in any textbook on probability and mathematical statistics.