QUANTITATIVE METHODS IN ECONOMICS
Vol. XVI, No. 2, 2015, pp. 152 – 158

A CONFIDENCE INTERVAL FOR PROPORTION IN FINITE POPULATION DIVIDED INTO TWO STRATA: A NUMERICAL STUDY

Wojciech Zieliński
Department of Econometrics and Statistics
Warsaw University of Life Sciences – SGGW
e-mail: wojciech_zielinski@sggw.pl

Abstract: Consider a finite population of N units. Let $\theta \in (0,1)$ denotes the fraction of units with a given property. The problem is in interval estimation of θ on the basis of a sample drawn due to the simple random sampling without replacement. Suppose, that the population is divided into two (disjoint) strata. In the paper the confidence interval for θ is proposed based on samples from two strata.

Keywords: confidence interval, sample size, fraction, finite population

The problem of the interval estimation of the fraction (proportion) θ is very old. The first solution was given by Clopper and Pearson [1934] and since then many authors deals with the problem. An exhaustive presentation of the problem along with the very rich literature may be found in the textbook by Koronacki and Mielniczuk [2009]. Presented solutions are valid for infinite populations. In many applications (economic, social, etc.) we deal with the finite population, so we are interested in interval estimation of θ in such finite populations. Remarks on differences in statistical inference in infinite and finite populations may be found in Bracha [1996].

Consider a population $U = \{u_1, \ldots, u_N\}$ containing the finite number N units. Let M denotes an unknown number of objects in population which has an interesting property. We are interested in an interval estimation of M, or equivalently, the fraction $\theta = \frac{M}{N}$. The sample of size n is drawn due to the simple random sampling without replacement. Let ξ be a random variable describing a number of objects with the property in the sample. On the basis of ξ we want to construct a confidence interval for θ at the confidence level δ.

The random variable ξ has the hypergeometric distribution [Johnson and Kotz 1969, Zieliński 2010]

$$P_{\theta,N,n}\{\xi = x\} = \frac{\binom{\theta N}{\mathrm{x}}\binom{(1-\theta)N}{\mathrm{n-x}}}{\binom{N}{\mathrm{n}}}$$

for integer x from the interval $\langle max\{0, n-(1-\theta)N\}, min\{n, \theta\, N\}\rangle$. Let $f_{\theta,N,n}(\cdot)$ be the probability distribution function, i.e.

$$f_{\theta,N,n}(x) = \begin{cases} P_{\theta,N,n}\{\xi = x\}, & for\ integer\ x \in \langle \max\{0, n-(1-\theta)N\}, \min\{n, \theta\, N\}\rangle \\ 0, & elsewhere \end{cases}$$

and let

$$F_{\theta,N,n}(x) = \sum_{t \le x} f_{\theta,N,n}(t)$$

be the cumulative distribution function of ξ. The CDF of θ may be written as

$$1 - \frac{\binom{n}{x+1}\binom{N-n}{\theta\, N-x-1}}{\binom{N}{\theta\, N}} \cdot \ {}_3F_2[\{1, x+1-\theta\, N, x+1-n\}, \{x+2, (1-\theta)N+x+2-n\}; 1],$$

where

$${}_3F_2[\{a_1, a_2, a_3\}, \{b_1, b_2\}; t] = \sum_{k=0}^{\infty} \left(\frac{(a_1)_k (a_2)_k (a_3)_k}{(b_1)_k (b_2)_k}\right)\left(\frac{t^k}{k!}\right)$$

and $(a)_k = a(a+1)\cdots(a+k-1)$.

A construction of the confidence interval at a confidence level δ for θ is based on the cumulative distribution function of ξ. If $\xi = x$ is observed then the ends

$$\theta_L = \theta_L(x-1, N, n, \delta_1) \text{ and } \theta_U = \theta_U(x, N, n, \delta_2)$$

of the confidence interval are the solutions of the two following equations

$$F_{\theta_L,N,n}(x-1) = \delta_1, F_{\theta_U,N,n}(x) = \delta_2.$$

The numbers δ_1 and δ_2 are such that $\delta_1 - \delta_2 = \delta$. In what follows we take $\delta_1 = (1+\delta)/2$ and $\delta_2 = (1-\delta)/2$. For $\xi = 0$ the left end is taken to be 0, and for $\xi = n$ the right end is taken to be 1. Analytic solution is unavailable. However, for given x, n and N, the confidence interval may be found numerically.

The hypergeometric distribution is analytically and numerically untractable. Hence different approximations are applied. There are at least two approximations commonly used in applications: Binomial and Normal. But using those approximations may lead to wrong conclusions [c.f Zieliński 2011]. So in what follows the exact distribution of ξ will be used.

Suppose that the population is divided into two strata: $U_1 = \{u_{11}, \ldots, u_{1N_1}\}$ and $U_2 = \{u_{21}, \ldots, u_{2N_2}\}$. Of course, $N_1 + N_2 = N$ and $U_1 \cap U_2 = \emptyset$. Let θ_1 and θ_2 be fractions of marked out units in the first and the second strata, respectively. The fraction of marked out units in the whole population equals

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

where $w_1 = N_1/N$ and $w_2 = N_2/N$.

It is known that for stratified populations it is better (sometimes) to estimate the proportion in the whole population using the information of stratification. Let n_1 and n_2 be the sizes of the samples drawn from the first and second strata due to the simple random sampling without replacement scheme, and let ξ_1 and ξ_2 be the random variables describing a number of "successes" in the first and second sample, respectively. The whole sample size is $n = n_1 + n_2$. The unbiased with the minimal variance estimator of θ is of the form

$$\hat{\theta}_s = w_1\hat{\theta}_1 + w_2\hat{\theta}_2,$$

where

$$\hat{\theta}_1 = \frac{\xi_1}{n_1} \quad \text{and} \quad \hat{\theta}_2 = \frac{\xi_2}{n_2}.$$

The variance of that estimator equals

$$var(\hat{\theta}_s) = w_1^2 \frac{\theta_1(1-\theta_1)}{n_1} \frac{N_1 - n_1}{N_1 - 1} + w_2^2 \frac{\theta_2(1-\theta_2)}{n_2} \frac{N_2 - n_2}{N_2 - 1},$$

while the variance of the estimator

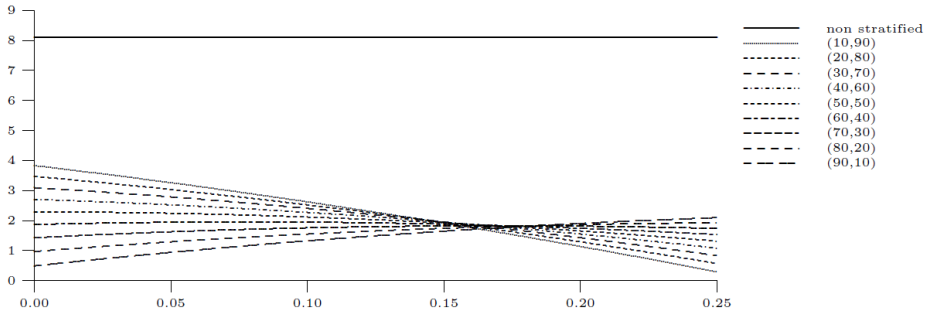
$$\hat{\theta} = \frac{\xi}{n}$$

of θ in non stratified population is

$$var(\hat{\theta}) = \frac{\theta \ (1-\theta \ )}{n} \frac{N \ - n}{N \ - 1}.$$

The comparison of those variances (see Figure 1) shows that for given θ and for all values of θ_1 (x-axis) and different allocations (n_1, n_2) the stratified estimator is better than non stratified ($N = 1000$, $N_1 = 400$, $N_2 = 600$, $n = 100$, $\theta = 0.1$). Of course, $\theta_2 = (\theta - w_1\theta_1)/w_2$.

Figure 1. Variances of $\hat{\theta}$ and $\hat{\theta}_s$ for $\theta = 0.1$



Source: own preparation

Now the question is, how to construct a confidence interval for θ on the basis of observed values x_1 and x_2 of r.v's ξ_1 and ξ_2 respectively. It may be expected that such confidence interval may be "better" than the confidence interval in the whole (non stratified) population. Let

$$\theta_L^1 = \theta_L^1(x_1 - 1, N_1, n_1, \gamma_{11}) \; and \; \theta_U^1 = \theta_U^1(x_1, N_1, n_1, \gamma_{21})$$

and

$$\theta_L^2 = \theta_L^2(x_2 - 1, N_2, n_2, \gamma_{12}) \; and \; \theta_U^2 = \theta_U^2(x_2, N_2, n_2, \gamma_{22})$$

be confidence intervals for θ_1 and θ_2 respectively. The confidence levels of those intervals are γ_1 and γ_2, i.e.

$$\gamma_{11} - \gamma_{21} = \gamma_1 \; and \; \gamma_{12} - \gamma_{22} = \gamma_2.$$

Consider the interval with the ends

$$\theta_L^s = w_1\theta_L^1 + w_2\theta_L^2 \; and \; \theta_U^s = w_1\theta_U^1 + w_2\theta_U^2.$$

The interval above may be considered as a confidence interval for θ constructed on the basis of two samples drawn from two strata.
The confidence level of the above interval equals

$$P_\theta\{\theta \in (\theta_L^s, \theta_U^s)\} = \frac{1}{H}\sum_{\theta_1=L}^{U}\sum_{x_1,x_2} P_{\theta_1}\{\xi_1 = x_1\}P_{\frac{\theta - w_1\theta_1}{w_2}}\{\xi_2 = x_2\}\mathbf{1}_{(w_1\theta_L^1(x_1)+w_2\theta_L^2(x_2), w_1\theta_U^1(x_1)+w_2\theta_U^2(x_2))}(\theta),$$

where

$$L = max\left\{0, \frac{\theta - w_2}{w_1}\right\}, \qquad U = min\left\{1, \frac{\theta}{w_1}\right\}, \qquad H = min\{1, \frac{\theta}{w_1}\} - max\{0, \theta - \frac{w_{2)}}{w_1}\} + \frac{1}{N_1}$$

and

$$\mathbf{1}_A(\theta) = \begin{cases} 1, & if \; \theta \in A \\ 0, & if \; \theta \notin A. \end{cases}$$

The expected length of the confidence level equals

$$d(\theta) = \sum_{x_1,x_2}(\theta_U^s - \theta_L^s)P_\theta\{\xi_1 = x_1, \xi_2 = x_2\}\mathbf{1}_{(\theta_L^s,\theta_U^s)}(\theta).$$

The main problem is to find γ_1 and γ_2 such that the confidence level of the confidence interval for θ is at least δ. The problem seems to be unsolvable analytically, so an appropriate numerical study was performed.

Numerical study

In the numerical study the following values were employed:

$$N = 1000, \quad N_1 = 400, \qquad N_2 = 600.$$

The overall sample size were taken $n = 100$. As the confidence level of the confidence interval for θ the value $\delta = 0.95$ was taken.

The numerical study had two aims. Firstly, we want to determine values of γ_1 and γ_2 such that the confidence level of the confidence interval for θ is δ . We assume that $\gamma_1 = \gamma_2 = \gamma$.

The second aim was to compare the lengths of the confidence intervals obtained for different allocations of the sample with the length of the confidence interval obtained for the non stratified population.

In Table 1 there are given confidence levels for different values of γ and different sample allocations. It may be seen that none of the proposed $\gamma's$ gives the prescribed confidence level δ.

Table 1. Confidence levels

(n_1,n_2) (10,90)	γ=0.85 0.94255	γ=0.86 0.96743	(n_1,n_2) (20,80)	γ=0.80 0.94493	γ=0.81 0.95373	(n_1,n_2) (30,70)	γ=0.80 0.94862	γ=0.81 0.95121
(n_1,n_2) (40,60)	γ=0.83 0.94359	γ=0.84 0.96172	(n_1,n_2) (50,50)	γ=0.81 0.94983	γ=0.82 0.95568	(n_1,n_2) (60,40)	γ=0.81 0.94539	γ=0.82 0.95059
(n_1,n_2) (70,30)	γ=0.83 0.94998	γ=0.84 0.95687	(n_1,n_2) (80,20)	γ=0.83 0.94532	γ=0.84 0.9519	(n_1,n_2) (90,10)	γ=0.83 0.93834	γ=0.84 0.95489

Source: the Author's calculations

Because of the discreetness of the r.v's ξ, no more accurate results are available. For example, for allocation (10,90) there exists $0.85 < \gamma^* < 0.86$ such that for $\gamma \leq \gamma^*$ the confidence level equals 0.94255 and equals 0.96743 otherwise. For length comparison we took the probability γ such that the confidence level is as near 0.95 as possible. Chosen values of γ for different allocations are given in Table 2.

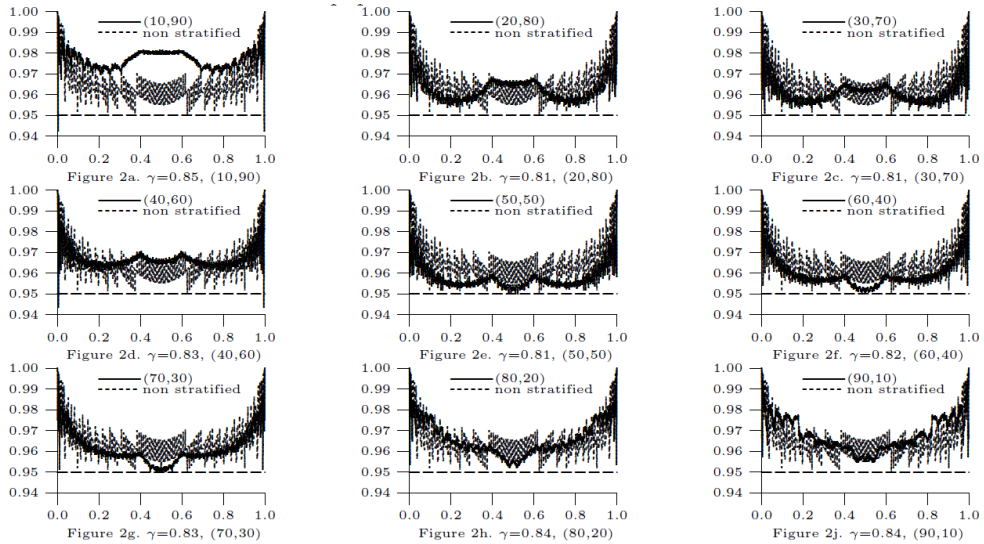
Table 2. Confidence levels

(n_1, n_2)	(10,90)	(20,80)	(30,70)	(40,60)	(50,50)	(60,40)	(70,30)	(80,20)	(90,10)
γ	0.85	0.81	0.81	0.83	0.81	0.82	0.83	0.84	0.84

Source: the Author's calculations

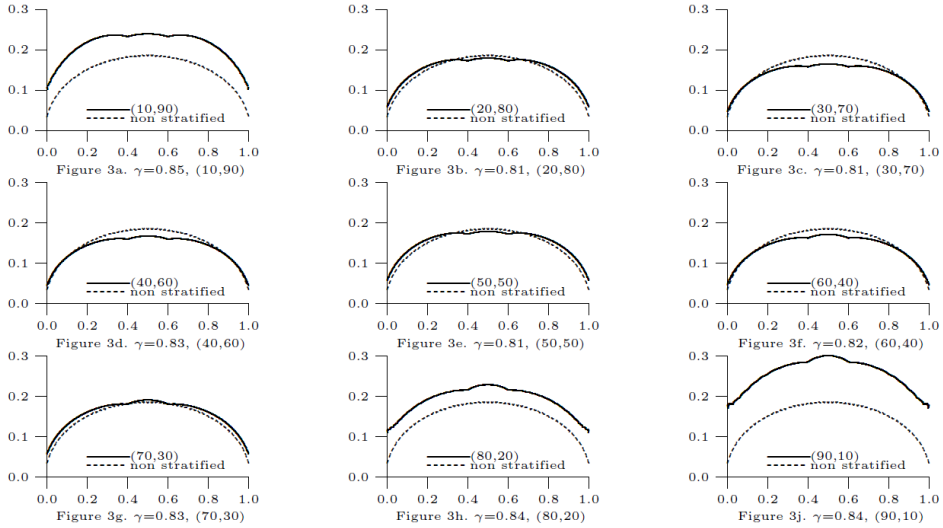
In the following figures there are presented confidence levels of the confidence intervals for stratified population as well as for the non stratified population.

Figure 2. Confidence levels of confidence intervals



Source: the Author's preparation

Figure 3. Comparison of confidence intervals



Source: the Author's preparation

In Figure 3 lengths of the proposed confidence interval are compared with the length of the non stratified confidence interval. It is seen that the use of the information of stratification gives worse results for $n_1 = 10, 80, 90$; comparable lengths for $n_1 = 20, 50, 70$ and shorter confidence interval otherwise.

Final remarks

Due to the Author knowledge confidence intervals for θ in stratified population were never considered, so the presented confidence interval is the first such proposition. There arises some very important questions with respect to the proposed confidence interval. The first one concerns of choosing γ_1 and γ_2: what values they should take on to obtain the prescribed confidence level of the confidence interval for θ in the whole population. The second question is of the optimal sample size and its allocation between two strata. The last but not least problem is the generalization of presented confidence interval for θ to the case of more than two strata.

REFERENCES

Bracha Cz. (1996) Teoretyczne podstawy metody reprezentacyjnej, PWN, Warszawa.

Johnson N. L., Kotz S. (1969) Discrete distributions: distributions in statistics, Houghton Mifflin Company, Boston.

Koronacki J., Mielniczuk J. (2009) Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT, Warszawa.

Neyman J. (1934) On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, Journal of the Royal Statistical Society, Vol. 97, No. 4, pp. 558-625.

Zieliński W. (2010) Estymacja wskaźnika struktury, Wydawnictwo SGGW, Warszawa.

Zieliński W. (2011) Comparison of confidence intervals for fraction in finite populations, Quantitative Methods in Economics, Vol. XII, pp. 177-182.