QUANTITATIVE METHODS IN ECONOMICS
Vol. XII, No. 1, 2011, pp. 177–182

COMPARISON OF CONFIDENCE INTERVALS FOR FRACTION IN FINITE POPULATIONS

Wojciech Zieliński
Department of Econometrics and Statistics
Warsaw University of Life Sciences – SGGW
e-mail: wojtek.zielinski@statystyka.info

Abstract. Consider a finite population. 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. In the paper three confidence intervals are compared: exact based on hypergeometric distribution and two other based on approximations to hypergeometric distribution: Binomial and Normal. It appeared that Binomial based confidence interval is too conservative while the Normal based one does not keep the prescribed confidence level.

Keywords: confidence interval, approximate confidence interval, fraction, finite population

INTRODUCTION

Consider a population $\{u_1, \ldots, u_N\}$ 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 $\theta = \dfrac{M}{N}$. The sample of size n is drawn due to the simple random sampling without replacement (*lpbz* to be short). Let ξ_{bz} be a random variable describing a number of objects with the property in the sample. Its distribution is hypergeometric (Bracha 1996, Zieliński 2010)

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

for integer x from the interval $\langle\max\{0,n-(1-\theta)N\},\min\{n,\theta N\}\rangle$. Denote by $f_\theta(\cdot)$ and $F_\theta(\cdot)$ the probability distribution function and cumulative distribution function of ξ_{bz}, respectively.

Note that

$$E_\theta\xi_{bz}=n\theta,\qquad D_\theta^2\xi_{bz}=\frac{N-n}{N-1}n\theta(1-\theta).$$

A construction of the confidence interval at a confidence level δ for θ is based on the cumulative distribution function of ξ_{bz}. If $\xi_{bz}=x$ is observed then the ends θ_L and θ_U of the confidence interval are the solutions of the two following equations

$$F_{\theta_L}(x)=\delta_1,\qquad F_{\theta_U}(x)=\delta_2.$$

The numbers δ_1 and δ_2 are such that $\delta_2-\delta_1=\delta$. In what follows we take $\delta_1=(1-\delta)/2$ and $\delta_2=(1+\delta)/2$. Analytic solution is unavailable. However, for given x, n and N, the confidence interval may be found numerically. In the Table 1 there are given exemplary confidence intervals for $N=1000$ units, sample size $n=20$, confidence level $\delta=0.95$ and $\delta_1=0.025$.

Table 1. Confidence intervals for θ

x	θ_L	θ_U	x	θ_L	θ_U	x	θ_L	θ_U
0	0.000	0.167	7	0.155	0.591	14	0.459	0.880
1	0.001	0.247	8	0.192	0.638	15	0.511	0.913
2	0.012	0.316	9	0.232	0.683	16	0.565	0.942
3	0.032	0.377	10	0.273	0.727	17	0.623	0.968
4	0.058	0.435	11	0.317	0.768	18	0.685	0.988
5	0.087	0.489	12	0.362	0.808	19	0.753	0.999
6	0.120	0.541	13	0.409	0.845	20	0.833	1.000

Source: own calculations

The real confidence level equals

$$conf_H = F_\theta(x_g) - F_\theta(x_d) = \sum_{x=x_d}^{x_g} f_\theta(x),$$

where

$$x_d = \max\left\{x : F_\theta(x) \le \frac{1-\delta}{2}\right\} \text{ and } x_g = \min\left\{x : F_\theta(x) \ge \frac{1+\delta}{2}\right\}.$$

Equivalently,

$$conf_H = \sum_{x=0}^{n} f_\theta(x)\mathbf{1}_{(\theta_L,\theta_U)}(\theta),$$

where $\mathbf{1}_A(a) = 1$ if $a \in A$ and $= 0$ elsewhere.

Since the population is finite, the number of admissible values of θ is also finite. For example, for $x = 1$ admissible values of θ are $0.001, 0.002, \ldots, 0.247$. It means, that the number of units with the investigated property is one of 1, 2, $\ldots$ or 247.

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.

BINOMIAL APPROXIMATION

The distribution of ξ_{bz}, for relatively small values of θ and large values of N may be approximated by Binomial distribution $Bin(n,\theta)$. As a rule of thumb $\theta < 0.1$ and $N \ge 60$ is sometimes used (Johnson and Kotz 1969).
The confidence interval for θ has the form (θ_L^B, θ_U^B), where

$$\theta_L^B = \beta^{-1}\left(\xi_{bz}, n-\xi_{bz}+1; \frac{1-\delta}{2}\right), \quad \theta_U^B = \beta^{-1}\left(\xi_{bz}+1, n-\xi_{bz}; \frac{1+\delta}{2}\right).$$

Here $\beta^{-1}(a,b;q)$ denotes the q-th quantile of the Beta distribution with parameters a,b (Clopper and Pearson 1934, Zieliński 2010). If $\xi_{bz} = 0$ then θ_L^B is taken 0. For $\xi_{bz} = n$, θ_U^B is taken 1. The true confidence level is equal to

$$conf_B = \sum_{x=0}^{n} f_\theta(x)\mathbf{1}_{(\theta_L^B,\theta_U^B)}(\theta).$$

NORMAL APPROXIMATION

The hypergeometric distribution may be approximated by a normal distribution with mean and variance equal to mean and variance of ξ_{bz}, i.e by the distribution $N(n\theta, \frac{N-n}{N-1} n\theta(1-\theta))$. To do so the rule of thumb $N\theta \geq 4$ is sometimes used (Johnson and Kotz 1969). The confidence interval for θ is obtained as a solution with respect to θ of the inequality

$$\left| \frac{\xi_{bz} - n\theta}{\sqrt{\frac{N-n}{N-1} n\theta(1-\theta)}} \right| \leq z_{\frac{1+\delta}{2}},$$

where z_q denotes the q-th quantile of the distribution $N(0,1)$. As a solution we obtain the confidence interval (θ_L^N, θ_U^N), where ($z = z_{\frac{1+\delta}{2}} \sqrt{\frac{n(N-n)}{N-1}}$)

$$\theta_L^N = \frac{z^2 + 2n\xi_{bz} - z\sqrt{z^2 + 4(n-\xi_{bz})\xi_{bz}}}{2(n^2 + z^2)}, \quad \theta_U^N = \frac{z^2 + 2n\xi_{bz} + z\sqrt{z^2 + 4(n-\xi_{bz})\xi_{bz}}}{2(n^2 + z^2)}.$$

The true confidence level is equal to

$$conf_N = \sum_{x=0}^{n} f_\theta(x) \mathbf{1}_{(\theta_L^N, \theta_U^N)}(\theta).$$

COMPARISON

Consider a population consisting $N = 1000$ units. From the population a sample of size $n = 20$ is drawn with respect to the *lpbz* scheme. The number ξ_{bz} of units with a given property is observed and confidence interval for the fraction θ of all such object in the population is constructed. The confidence level $\delta = 0.95$ is assumed.

In the Table 2 there are given confidence limits calculated on the basis of the exact distribution of ξ_{bz} as well as confidence limits based on two approximations.

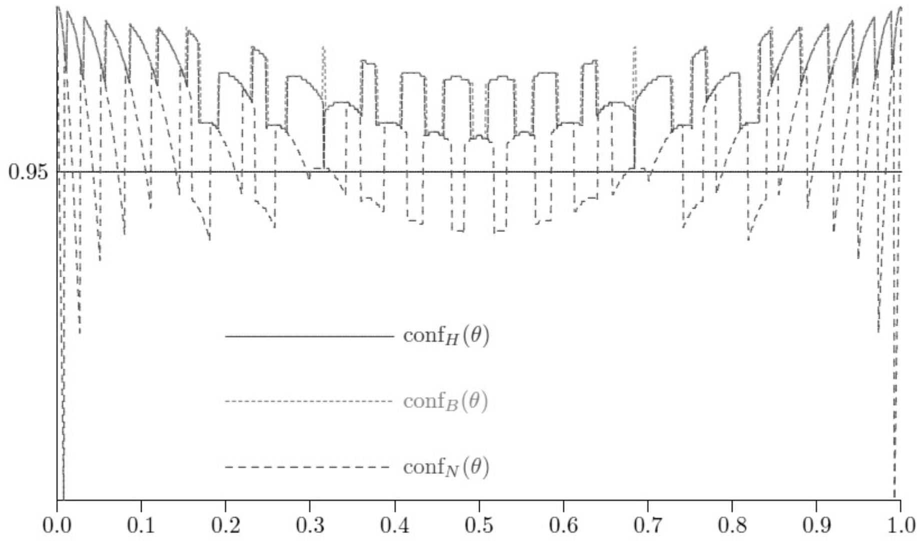
Table 2. Comparison of confidence intervals

	Hypergeometric		Binomial		Normal	
x	θ_L	θ_U	θ_L^B	θ_U^B	θ_L^N	θ_U^N
0	0.000	0.167	0.000	0.169	0.000	0.159
1	0.001	0.247	0.001	0.249	0.009	0.234
2	0.012	0.315	0.012	0.317	0.028	0.299
3	0.032	0.377	0.032	0.379	0.052	0.359
4	0.058	0.435	0.057	0.437	0.081	0.414
5	0.087	0.489	0.086	0.492	0.112	0.467
6	0.120	0.541	0.118	0.543	0.146	0.517
7	0.155	0.591	0.153	0.593	0.182	0.566
8	0.192	0.638	0.191	0.640	0.220	0.612
9	0.232	0.683	0.230	0.685	0.259	0.657
10	0.273	0.727	0.271	0.729	0.300	0.700
11	0.317	0.768	0.315	0.770	0.343	0.741
12	0.362	0.808	0.360	0.809	0.388	0.780
13	0.409	0.845	0.407	0.847	0.434	0.818
14	0.459	0.880	0.457	0.882	0.483	0.854
15	0.511	0.913	0.508	0.914	0.533	0.888
16	0.565	0.942	0.563	0.943	0.586	0.919
17	0.623	0.968	0.621	0.968	0.641	0.948
18	0.685	0.988	0.683	0.988	0.701	0.972
19	0.753	0.999	0.751	0.999	0.766	0.991
20	0.833	1.000	0.831	1.000	0.841	1.000

Source: own calculations

In the Figure there are shown real confidence levels of above confidence intervals. It may be noted that the Binomial based confidence interval is too conservative: its length is greater than the hypergeometric one and its confidence level is higher. The Normal approximation based confidence interval does not keep the prescribed confidence level (here 0.95), so this interval should not be used in fraction estimation.

Figure 1.



Source: own preparation

REFERENCES

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

Clopper C. J., Pearson E. S. (1934) The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial, Biometrika 26, 404-413.

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

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