QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 309 – 315

CONFIDENCE INTERVALS FOR FRACTION IN FINITE POPULATIONS: MINIMAL SAMPLE SIZE

Wojciech Zieliński
Department of Econometrics and Statistics
Warsaw University of Life Sciences – SGGW
Department of the Prevention of Environmental Hazards and Allergology
Medical University of Warsaw
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. It is of interest to obtain confidence intervals of a prescribed length. In the paper the minimal sample size which guarantees the length to not exceed the given value is calculated.

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

Consider a population $\{u_1, \dots, u_N\}$ 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 $\theta = M/N$. 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. On the basis of ξ_{bz} we want to construct a confidence interval for θ at the confidence level δ. The main problem is to find minimal sample size n such that the expected length of the confidence interval is smaller than the given number $\varepsilon > 0$.
The random variable ξ_{bz} has the hypergeometric distribution [Johnson and Kotz 1969, Zieliński 2010]

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

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

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

and let

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

be the cumulative distribution function of ξ_{bz}.
The CDF of ξ_{bz} may be written as

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

where

$${}_3F_2[\{1, x+1-\theta\, N, x+1-n\},\{x+2,(1-\theta)N+x+2-n\};1]=\sum_{k=0}^{\infty}\frac{(a_1)_k(a_2)_k(a_3)_k}{(b_1)_k(b_2)_k}\frac{t^k}{k!}$$

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_{bz}. If $\xi_{bz}=x$ is observed then the ends $\theta_L=\theta_L(x,N,n,\delta_1)$ and $\theta_U=\theta_U(x,N,n,\delta_2)$ of the confidence interval are the solutions of the two following equations

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

The numbers δ_1 and δ_2 are such that $\delta_2-\delta_1=\delta$. In what follows we take $\delta_1=\frac{1-\delta}{2}$ and $\delta_2=\frac{1+\delta}{2}$. 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 for $N=1000$, $n=20$, $\delta=0.95$

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 study

The real confidence level equals

$$conf_{N,n,\delta}(\theta) = \sum_{x=x_d}^{x_g} f_{\{\theta,N,n\}}(x)\,,$$

where

$$x_d = F^{-1}_{\{\theta,N,n\}}\left(1-\frac{\delta}{2}\right) \text{ and } x_g = F^{-1}_{\{\theta,N,n\}}\left(1+\frac{\delta}{2}\right).$$

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. Consider the length of the confidence interval

$$d(\xi_{bz}, N, n, \delta) = \theta_U(\xi_{bz}, N, n, \frac{1+\delta}{2}) - \theta_L(\xi_{bz}, N, n, \frac{1-\delta}{2}).$$

It is easy to note, that the length depends among others on the population size N and the sample size n.

Let $\varepsilon > 0$ be a given number. We are going to find minimal sample size n (for a given population size N) such that the length of the confidence interval is smaller than ε. More precisely, we are going to find minimal sample size n such that the expected length of confidence interval covering θ is smaller than ε, i.e.

$$L(\theta, N, n) = E_{\theta,N,n}\left(d(\xi_{bz}, N, n, \delta)1_{\left(\theta_L\left(\xi_{bz},N,n,\frac{1-\delta}{2}\right),\theta_U\left(\xi_{bz},N,n,\frac{1+\delta}{2}\right)\right)}(\theta)\right) \leq \varepsilon,$$

where

$$1_A(z) = \begin{cases} 1, & z \in A \\ 0, & z \notin A \end{cases}$$

Note that the expected length may be written as

$$L(\theta, N, n) = \sum_{x=x_d}^{x_g} d(x, N, n, \delta) f_{\theta,N,n}(x).$$

For given θ, N and ε the inequality

$$L(\theta, N, n) \leq \varepsilon$$

may be solved numerically. Exemplary solutions for $\theta = 0.05$ and $\varepsilon = 0.02$ are given in the Table 2.

Table 2. Minimal sample sizes for $\theta = 0.05$, $\varepsilon = 0.02$, $\delta = 0.95$

N	n_{min}	conf. level	Length	N	n_{min}	conf. level	length
500	410	0.970508	0.019380	5500	1352	0.956080	0.019994
1000	661	0.954932	0.019901	6000	1369	0.952244	0.019989
1500	836	0.957584	0.019865	6500	1401	0.955592	0.019990
2000	966	0.960184	0.019949	7000	1416	0.952569	0.019998
2500	1061	0.958530	0.019992	7500	1443	0.956406	0.019999
3000	1136	0.953237	0.019902	8000	1461	0.953642	0.019963
3500	1188	0.951010	0.019947	8500	1465	0.952202	0.019993
4000	1238	0.950565	0.019962	9000	1489	0.956288	0.019999
4500	1286	0.951242	0.019900	9500	1497	0.955144	0.019982
5000	1316	0.953628	0.019997	10000	1511	0.953024	0.019932

Source: own study

Note that the expected length is smaller than given ε. To obtain the expected length equal exactly ε a randomization is needed. This randomization is between sample sizes. Note that

$$L(\theta, N, n) \leq \varepsilon \leq L(\theta, N, n-1).$$

Let us choose in random the sample size:

$$sample\ size = \begin{cases} n-1, & with\ probability\ \gamma, \\ n, & with\ probability\ 1-\gamma, \end{cases}$$

where $\gamma = \frac{\varepsilon - L(\theta,N,n)}{L(\theta,N,n-1) - L(\theta,N,n)}$.

The expected length of the confidence interval after such randomization equals

$$\gamma\, L(\theta, N, n-1) + (1-\gamma) L(\theta, N, n) = \varepsilon.$$

For example, if $N = 1000$, $\varepsilon = 0.02$, $\delta = 0.05$ and $\theta = 0.05$ then

$$L(0.05,1000,661) = 0.019901\ and\ \ L(0.05,1000,660) = 0.020133.$$

Choosing $\gamma = 0.427158$ and applying the rule

$$sample\ size = \begin{cases} 660, & with\ probability\ 0.427158, \\ 661, & with\ probability\ 0.572842, \end{cases}$$

we obtain the confidence interval with the mean length equal to the prescribed accuracy.

In the Table 3 are given randomized minimum sample sizes. In the last column of the Table the probability γ is given.

Table 3. Randomized minimal sample sizes for $\theta = 0.05$, $\varepsilon = 0.02$, $\delta = 0.95$.

N	n_{min}	conf. level	length	$n_{min} - 1$	conf. level	length	γ
500	410	0.970508	0.019380	409	0.986565	0.020130	0.826673
1000	661	0.954932	0.019901	660	0.967684	0.020133	0.427158
1500	836	0.957584	0.019865	835	0.968619	0.020139	0.492457
2000	966	0.960184	0.019949	965	0.960078	0.020030	0.630721
2500	1061	0.958530	0.019992	1060	0.96769	0.020188	0.038677
3000	1136	0.953237	0.019902	1135	0.96203	0.020066	0.599829
3500	1188	0.951010	0.019947	1187	0.959462	0.020158	0.250663
4000	1238	0.950565	0.019962	1237	0.958750	0.020112	0.253571
4500	1286	0.951242	0.019900	1285	0.959274	0.020063	0.612891
5000	1316	0.953628	0.019997	1315	0.953633	0.020025	0.123186
5500	1352	0.956080	0.019994	1351	0.956157	0.020012	0.337816
6000	1369	0.952244	0.019989	1368	0.952291	0.020020	0.360976
6500	1401	0.955592	0.019990	1400	0.955660	0.020008	0.551318
7000	1416	0.952569	0.019998	1415	0.952588	0.020015	0.124424
7500	1443	0.956406	0.019999	1442	0.956399	0.020024	0.047004
8000	1461	0.953642	0.019963	1460	0.960810	0.020094	0.281286
8500	1465	0.952202	0.019993	1464	0.952217	0.020001	0.921897
9000	1489	0.956288	0.019999	1488	0.956253	0.020003	0.232190
9500	1497	0.955144	0.019982	1496	0.955181	0.020014	0.554152
10000	1511	0.953024	0.019932	1510	0.960018	0.020093	0.420791

Source: own study

Normal approximation. The hypergeometric distribution is analytically untractable, so in many applications the normal approximation of the distribution is used. The hypergeometric distribution, by Central Limit Theorem, 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 [Bracha 1996], [Karliński 2003], [Zieliński 2010].

In practice two approaches are met. In the first one the confidence interval is obtained as a solution of

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

where z_q denotes the q-th quantile of the distribution $N(0,1)$. The confidence interval takes on the form

$$\theta_L^N = \left(\frac{\xi_{bz}}{n}\right) - e, \theta_U^N = \left(\frac{\xi_{bz}}{n}\right) + e$$

and

$$e^2 = \frac{z_{\frac{1+\delta}{2}}^2}{n^3} \xi_{bz}(n - \xi_{bz}) \left(\frac{N-n}{N-1}\right).$$

In the second approach the confidence interval is obtained as a solution of

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

The confidence interval ($z = z_{\frac{1+\delta}{2}}\sqrt{\frac{n(N-n)}{N-1}}$) has the form

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

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

Table 4. First normal approximation of sample sizes for $\theta = 0.05$, $\varepsilon = 0.02$, $\delta = 0.95$.

N	n_{min}	conf. level	length	N	n_{min}	conf. level	length
500	385	0.950162	0.019979	5500	1267	0.944224	0.019930
1000	624	0.949319	0.019965	6000	1294	0.948367	0.020000
1500	786	0.942629	0.019891	6500	1317	0.944391	0.019899
2000	906	0.949799	0.019976	7000	1320	0.941959	0.019994
2500	989	0.939873	0.019903	7500	1349	0.946901	0.019989
3000	1064	0.946046	0.019966	8000	1368	0.943899	0.019895
3500	1116	0.944653	0.019993	8500	1370	0.942252	0.019959
4000	1166	0.944888	0.019952	9000	1394	0.947670	0.019984
4500	1204	0.946749	0.019986	9500	1400	0.946198	0.019989
5000	1243	0.949173	0.019993	10000	1419	0.943788	0.019869

Source: own study

Table 5. Second normal approximation of sample sizes for $\theta = 0.05$, $\varepsilon = 0.02$, $\delta = 0.95$.

N	n_{min}	conf. level	length	N	n_{min}	conf. level	length
500	384	0.948488	0.019988	5500	1261	0.939422	0.019866
1000	619	0.936714	0.019932	6000	1286	0.942464	0.019930
1500	790	0.935830	0.019619	6500	1310	0.939736	0.019843
2000	901	0.946060	0.019987	7000	1329	0.947937	0.019996
2500	989	0.934991	0.019777	7500	1336	0.941714	0.019980
3000	1063	0.939874	0.019824	8000	1360	0.939704	0.019852
3500	1113	0.939006	0.019889	8500	1361	0.938450	0.019935
4000	1162	0.939416	0.019860	9000	1385	0.942431	0.019925
4500	1195	0.943374	0.019998	9500	1386	0.941560	0.019991
5000	1236	0.942720	0.019908	10000	1409	0.939789	0.019844

Source: own study

Comparing minimal sample sizes obtained in the exact solution (Table 2) to normal approximations (Tables 4 and 5) it is seen that application of approximate

confidence intervals needs smaller sample sizes. But the confidence level of approximate confidence intervals does not keep the nominal confidence level, so the risk of wrong conclusions is too high (greater than nominal).

In some applications the Binomial approximation to the hypergeometric distribution is applied. Comparison of all approximations may be found in [Zieliński 2011].

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.

Karliński W. (2003) Nowe techniki w kontroli wykonania budżetów państwa, Kontrola Państwowa 5/2003, 101-124.

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

Zieliński W. (2011) Comparison of confidence intervals for fraction in finite populations, Metody Ilościowe w Badaniach Ekonomicznych XII, 177-182.