STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 115—126

ROBUSTNESS OF THE CONFIDENCE INTERVAL FOR AT-RISK-OF-POVERTY RATE

Wojciech Zieliński[1]

ABSTRACT

Zieliński (2009) constructed a nonparametric interval for *At-Risk-of-Poverty Rate*. It appeared that the confidence level of the interval depends on the underlying distribution of the income. For some distributions (e.g. lognormal, gamma, Pareto) the confidence level may be smaller than the nominal one. The question is, what is the largest deviance from the nominal level?

In the paper, a more general problem is considered, i.e. the problem of robustness of the confidence level of the confidence interval for binomial probability. The worst distribution is derived as well as the smallest true confidence level is calculated. Some asymptotic remarks (sample size tends to infinity) are also given.

Key words: ARPR, binomial distribution, confidence interval, robustness.

1. Confidence interval for ARPR

In the European Commission Eurostat document Doc. IPSE/65/04/EN page 11, the „*at-risk-of-poverty rate*" (ARPR) is defined as the percentage of population with the equivalised disposable income below 60% of calculated median value, i.e. ARPR is defined as

$$ARPR = F(0.6 \cdot Q(0.5)),$$

where F denotes the distribution of the population income ($Q(0.5)$ stands for the median of the distribution F).

The natural estimator of ARPR is as follows. Let $X_1, \ldots, X_n$ be a sample of disposable incomes of randomly drawn n persons. Let $X_{1:n} \le \cdots \le X_{n:n}$ denote the ordered sample. As an estimator of the population median we take $X_{M:n}$,

[1] Department of Econometrics and Statistics, Warsaw University of Life Sciences, Nowoursynowska 159, 02-776 Warszawa. E-mail: wojtek.zielinski@statystyka.info.

where $M = \lfloor n/2 \rfloor + 1$ ($\lfloor a \rfloor$ is the greatest integer not greater than a). Let ξ be the number of observations not greater than $0.6 \cdot X_{M:n}$:

$$\xi = \#\{X_i \leq 0.6 \cdot X_{M:n}\}.$$

Here, $\# S$ denotes the cardinality of the set S. The estimator $\widehat{ARPR}$ is defined:

$$\widehat{ARPR} = \frac{\xi}{n}.$$

The properties of the above estimator depends strongly on the distribution F of population income. Zieliński (2009) showed that the distribution of ξ is almost binomial with parameters $M-1$ and $2 \cdot ARPR$. Zieliński (2006) showed that the estimator is almost unbiased, i.e.

$$E_F \widehat{ARPR} \approx ARPR$$

for all continuous F. He also calculated its variance:

$$D_F^2 \widehat{ARPR} \approx \frac{1}{n} ARPR(1 - ARPR).$$

The nonparametric interval for *ARPR* at a confidence level $\gamma \in (0,1)$ is based on the random variable ξ (Zieliński 2009):

$$\left(0.5B^{-1}\left(\xi, M-\xi+1; \frac{1-\gamma}{2}\right); 0.5B^{-1}\left(\xi+1, M-\xi; \frac{1-\gamma}{2}\right)\right), (*)$$

where $B^{-1}(a,b;\delta)$ is the δ quantile of beta distribution with parameters (a,b).

The above confidence interval is a special case of the Clopper and Pearson (1934) confidence interval for binomial proportion. In the binomial statistical model

$$\left(\{0,1,\ldots,n\}, \{Bin(n,q), 0 < q < 1\}\right),$$

where $Bin(n,q)$ denotes binomial distribution, the confidence interval for q is of the form

$$\left(\beta^{-1}\left(\eta, n-\eta+1; \frac{1-\gamma}{2}\right); \beta^{-1}\left(\eta+1, n-\eta; \frac{1+\gamma}{2}\right)\right).$$

Here, η denotes the random variable distributed as $Bin(n,q)$.

2. Robustness of the confidence level

Suppose now that η is not distributed as $Bin(n,q)$, but its distribution is given by the cdf function F from a class F_q such that $F \in F_q$ iff

- F is discrete: $P_F(i) = p_i \geq 0$ for $i \in \{0,1,\ldots,n\}$;
- $\sum_{i=0}^{n} p_i = 1$;
- $\sum_{i=0}^{n} i p_i = nq$;
- $\sum_{i=0}^{n} i^2 p_i = nq(1-q) + (nq)^2$.

In the class F_q there are distributions with the same expected value and variance as binomial $Bin(n,q)$. We want to calculate

$$\inf_{q\in[0,1]} \inf_{F\in F_q} P_F \left\{ q \in \left(\beta^{-1}\left(\eta, n-\eta+1; \frac{1-\gamma}{2} \right); \beta^{-1}\left(\eta+1, n-\eta; \frac{1+\gamma}{2} \right) \right) \right\},$$

i.e.

$$\inf_{q\in[0,1]} \inf_{F\in F_q} \sum_{i=\lceil x_1(q) \rceil}^{\lceil x_2(q)+1 \rceil} p_i,$$

where $\lceil x_1(q) \rceil$ and $\lceil x_2(q) \rceil$ are such that

$$\beta(x_1(q)+1, n-x_1(q); q) = \frac{1+\gamma}{2} \quad \text{and} \quad \beta(x_2(q), n-x_2(q)+1; q) = \frac{1-\gamma}{2}.$$

For any given q (because of the symmetry of the problem, we consider 0<q<0.5), we have a linear programming problem with respect to $p_0, p_1, \ldots, p_n$:

$$\begin{cases} \sum_{i=\lceil x_1(q) \rceil}^{\lceil x_2(q)+1 \rceil} p_i = \min! \\ \sum_{i=0}^{n} p_i = 1 \\ \sum_{i=0}^{n} i p_i = nq \\ \sum_{i=0}^{n} i^2 p_i = nq(1-q) + (nq)^2 \\ 0 \leq p_i \leq 1, \ i = 0,1,\ldots,n \end{cases}$$

The solution of this problem is a one of vertices of the appropriate simplex, i.e. the point such that exactly three probabilities are not zero and others equal zero. Each vertex is a solution of the system of linear equations

$$\begin{bmatrix} 1 & 1 & 1 \\ a & b & c \\ a^2 & b^2 & c^2 \end{bmatrix}\begin{bmatrix} p_a \\ p_b \\ p_c \end{bmatrix}=\begin{bmatrix} 1 \\ w \\ z \end{bmatrix},$$

where a, b, c are integers such that $0 \le a < b < c \le n$, $w = nq$, $z = nq(1-q)+(nq)^2$. The solution is

$$p_a(a,b,c) = -\frac{-bc+(b+c)w-z}{(a-b)(a-c)},$$

$$p_b(a,b,c) = -\frac{ac-(a+c)w+z}{(a-b)(b-c)},$$

$$p_c(a,b,c) = -\frac{-ab+(a+b)w-z}{(a-c)(b-c)},$$

Among all solutions we have to choose one such that $0 < p_a(a,b,c), p_b(a,b,c), p_c(a,b,c) < 1$. It is enough to find a, b, c such that $p_a(a,b,c), p_b(a,b,c), p_c(a,b,c) > 0$. Solving the appropriate system of inequalities we obtain:

$$\begin{cases} 0 \le a < b \le (n-1)q \;\&\; nq\left(1+\dfrac{1-q}{nq-a}\right) < c < nq\left(1+\dfrac{1-q}{nq-b}\right) \\ \text{or} \\ 0 \le a \le (n-1)q \;\&\; (n-1)q < b \le nq\left(1+\dfrac{1-q}{nq-a}\right) \;\&\; nq\left(1+\dfrac{1-q}{nq-a}\right) < c \le n \end{cases}$$

(*)

We are looking for values a, b, c satisfying (*), such that, if F is concentrated in those points, then

$$P_F\left\{q \in \left(\beta^{-1}\left(\eta, n-\eta+1; \frac{1-\gamma}{2}\right); \beta^{-1}\left(\eta+1, n-\eta; \frac{1+\gamma}{2}\right)\right)\right\},$$

(P)

is the smallest. Note that

$$(P)=\begin{cases} 1 \quad \textit{if } \lfloor x_1(q)\rfloor \le a < b < c \le \lfloor x_2(q)\rfloor+1 \quad A \\ \mathrm{p_a}(\mathrm{a,\ b,\ c}) + \mathrm{p_b}(\mathrm{a,\ b,\ c}), \quad \textit{if } \lfloor x_1(q)\rfloor \le a < b \le \lfloor x_2(q)\rfloor+1 < c \quad B \\ \mathrm{p_a}(\mathrm{a,\ b,\ c}), \quad \textit{if } \lfloor x_1(q)\rfloor \le a \le \lfloor x_2(q)\rfloor+1 < b < c, \quad C \\ \mathrm{p_b}(\mathrm{a,\ b,\ c}) + \mathrm{p_c}(\mathrm{a,\ b,\ c}), \quad \textit{if } a < \lfloor x_1(q)\rfloor \le b < c \le \lfloor x_2(q)\rfloor+1, \quad D \\ \mathrm{p_b}(\mathrm{a,\ b,\ c}), \quad \textit{if } a < \lfloor x_1(q)\rfloor \le b \le \lfloor x_2(q)\rfloor+1 < c, \quad E \\ \mathrm{p_c}(\mathrm{a,\ b,\ c}), \quad \textit{if } a < b < \lfloor x_1(q)\rfloor \le c \le \lfloor x_2(q)\rfloor+1, \quad F \\ 0, \quad \textit{elswhere} \end{cases}$$

Case (A) is not interesting. Case (G) does not hold because $\lfloor x_1(q)\rfloor < nq < \lfloor x_2(q)\rfloor+1$ and hence at least one inequality holds

$$\lfloor x_1(q)\rfloor < a < \lfloor x_2(q)\rfloor+1, \quad \lfloor x_1(q)\rfloor < b < \lfloor x_2(q)\rfloor+1, \quad \lfloor x_1(q)\rfloor < c < \lfloor x_2(q)\rfloor+1.$$

It easy to see that the minimum of (P) is achieved in case (E), i.e. for *a, b, c* such that

$$a < \lfloor x_1(q) \le b \le \lfloor x_2(q)\rfloor+1 < c$$

and $q\in(0,0.5)$ such that

$$\lfloor x_1(q)\rfloor \ge 1.$$

So we consider probabilities q from the interval $(q(n),0.5)$, where $q(n)$ is the solution of

$$\beta(2,n-1;q(n))=\frac{1+\gamma}{2} \quad \text{i.e.} \quad \frac{1}{B(2,n-1)}\int_0^{q(n)} u(1-u)^{n-2}\,du=\frac{1+\gamma}{2}.$$

Exemplary values of $q(n)$ are given in the Table 1.

Table 1.

n	10	50	100	500	1000
q(n)	0.482497	0.108542	0.054997	0.011115	0.005564

The distribution minimizing (P) for a given q is given in theorem 1.

Theorem 1. Let $q \in (q(n), 0.5)$. Then

$$P_F\left\{q \in \left(\beta^{-1}\left(\eta, n-\eta+1; \frac{1-\gamma}{2}\right); \beta^{-1}\left(\eta+1, n-\eta; \frac{1+\gamma}{2}\right)\right)\right\}$$

reaches minimum at the distribution F concentrated in points

$$a = \lfloor x_1(q) \rfloor - 1,\ b = \frac{\lfloor x_1(q) \rfloor + \lfloor x_2(q) \rfloor + 1}{2},\ c = \lfloor x_2(q) \rfloor + 2$$

Proof. Proof of the theorem is given in the Appendix.

For a given q the distribution F minimizing probability (P) (or maximizing probability $prob(a,c)$) may be found. This distribution is concentrated in points

$$a = \max\left\{d \in \mathbf{N} : \beta^{-1}\left(d+1, n-d; \frac{1+\gamma}{2}\right) \le q\right\}, c = \min\left\{d \in \mathbf{N} : \beta^{-1}\left(d, n-d+1; \frac{1-\gamma}{2}\right) \ge q\right\}$$

and $b = \frac{a+c}{2}$. It is clear that (P) does not depend only on F but also on q. Of course, it is because of discreetness of considered distributions. For example, let $n = 20$, $a = 0$ and $c = 8$. Then, $b = 4$, $\lfloor x_1(q) \rfloor = 1$ and $\lfloor x_2(q) \rfloor + 1 = 7$ and for ($\gamma = 0.95$)

$$q \in \left(\beta^{-1}\left(1, 20; \frac{1+\gamma}{2}\right), \beta^{-1}\left(7, 13; \frac{1-\gamma}{2}\right)\right) = (0.16843, 0.19119)$$

confidence level of the confidence interval equals

$$\sum_{i=1}^{7} p_i = p_4(0,4,8) = q\frac{35-95q}{4}.$$

Values of the above probability are calculated in the Table 2.

Table 2.

q	p_4(0,4,8)	prob(0,8)	q	p_4(0,4,8)	prob(0,8)
0.169	0.80043	0.19957	0.181	0.80568	0.19432
0.170	0.80113	0.19888	0.182	0.80581	0.19420
0.171	0.80178	0.19822	0.183	0.80589	0.19411
0.172	0.80238	0.19762	0.184	0.80592	0.19408
0.173	0.80294	0.19706	0.185	0.80591	0.19409
0.175	0.80391	0.19609	0.186	0.80585	0.19416
0.176	0.80432	0.19568	0.187	0.80574	0.19426
0.177	0.80469	0.19531	0.188	0.80558	0.19442
0.178	0.80501	0.19500	0.189	0.80538	0.19462
0.179	0.80528	0.19472	0.190	0.80513	0.19488
0.180	0.80550	0.19450	0.191	0.80483	0.19517

It may be seen that minimal confidence level, i.e. maximal $prob(a,c)$, is attained at the $q = \beta^{-1}\left(1, 20; \frac{1+\gamma}{2}\right)$. This fact is generalized in the theorem 2.

Theorem 2. Let a be given and let

$$q_1 = \beta^{-1}\left(a+1, n-a; \frac{1+\gamma}{2}\right) \quad \text{and} \quad q_2 = \beta^{-1}\left(c, n-c+1; \frac{1-\gamma}{2}\right),$$

where $c = \min\left\{d \in \mathbf{N} : \beta^{-1}\left(d, n-d+1; \frac{1-\gamma}{2}\right) \geq q_1\right\}$. Then, $prob(a,c)$ is reached for q_1.

Proof. See Appendix

In order to find maximum of the probability $prob(a,c)$ it must be calculated for

$$a = 0, 1, \ldots, \frac{n-1}{2},\ \beta^{-1}\left(a+1, n-a; \frac{1+\gamma}{2}\right) \text{and c given in Theorem 2.}$$

For given n and γ it may be found by numerical solution. In the Table 3 there are calculated the minimal confidence level (*min*), nominal (*nom*) confidence level (i.e. in binomial model) and $percent = min / nom$.

Table 3.

n	min	nom	percent		n	min	nom	percent
10	0.8118	0.9625	0.8434		100	0.7645	0.9504	0.8044
15	0.7935	0.9517	0.8338		200	0.7573	0.9500	0.7972
20	0.7988	0.9580	0.8338		300	0.7547	0.9502	0.7942
30	0.7833	0.9506	0.8240		400	0.7540	0.9503	0.7935
40	0.7756	0.9503	0.8162		500	0.7534	0.9500	0.7930
50	0.7777	0.9533	0.8157		600	0.7532	0.9503	0.7926
60	0.7699	0.9503	0.8101		700	0.7531	0.9501	0.7926
70	0.7691	0.9508	0.8089		800	0.7524	0.9502	0.7918
80	0.7687	0.9513	0.8080		900	0.7520	0.9502	0.7914
90	0.7671	0.9512	0.8065		1000	0.7518	0.9501	0.7913

For large n, the binomial distribution for $q \in (q(n), 0.5)$ is almost normal $N(nq, nq(1-q))$ and

$$a, c \approx nq \pm u\sqrt{nq(1-q)},$$

where u is the $\frac{1+\gamma}{2}$ quantile of $N(0,1)$. It is easy to check that

$$prob(a,c) = \frac{1}{u^2}$$

and asymptotic minimal confidence level is $1 - \frac{1}{u^2}$. For example, if $\gamma = 0.95$ then $u = 1.96$, so asymptotic minimal confidence level is 0.73969.

3. Conclusions

The confidence level of the Clopper-Pearson confidence interval for probability in binomial model depends on the distribution from the class F_q. The minimal probability of coverage was calculated. It appeared that the most unfavourable distribution is a three point one.

From practical point of view it may be interesting to consider another family of distributions, namely a class $F_q^\varepsilon \subset F_q$: a distribution $F \in F_q^\varepsilon$ iff $|F(x) - q(x)| < \varepsilon$ for all $x = 0, 1, \ldots, n$. Here, ε is a given number. Such considerations are in progress.

4. Appendix

The proof of the Theorem 1 is given in two following lemmas.

Lemma 1. For given *a, c* satisfying $(*)$, the function $b \to p_b(a,b,c)$ achieves minimum for $b = \frac{a+c}{2}$.

Proof. Assume that b is real. Then

$$\frac{\partial p_b(a,b,c)}{\partial b} = \frac{(a-2b+c)(a(c-w)-cw+z)}{(a-b)^2(b-c)^2}.$$

For *a, c* from $(*)$

$$a(c-w)-cw+z = (a-w)c+z-aw < 0.$$

Hence $\frac{\partial p_b(a,b,c)}{\partial b} < 0$ for $b < \frac{a+c}{2}$ and $\frac{\partial p_b(a,b,c)}{\partial b} > 0$ for $b > \frac{a+c}{2}$.

Lemma 2. Let $b = \frac{a+c}{2}$. Then

$\max_{a,c}(p_a(a,b,c)+p_c(a,b,c)) = p_a(a^*,b,c^*)+p_c(a^*,b,c^*)$ for

$a^* = \max\{a < \frac{nw-z}{n-w}$ and $c^* = \min\{c > \frac{aw-z}{a-w}$.

Proof. Assume for a while that a and c are real numbers. We have

$$prob(a,c) = p_a(a,b,c)+p_c(a,b,c) = \frac{a^2+c^2+2a(c-2w)-4cw+4z}{(c-a)^2} = \frac{(a+c-2w)^2+4(z-w^2)}{(c-a)^2}.$$

The proof of the Lemma will be given in two steps

• for all $c > \frac{aw-z}{a-w}$ function $prob(a,c)$ is increasing with respect to a;

∗ for all $a < \frac{nw-z}{n-w}$ function $prob(a,c)$ for $c > \frac{aw-z}{a-w}$ achieves maximum for the smallest c.

Proof of •: we will show that

$$\frac{\partial}{\partial a} prob(a,c) = -\frac{4(c^2 + a(c-w) - 3cw + 2z)}{(a-c)^3} > 0 \quad \text{for} \quad 0 < a < \frac{nw-z}{n-w}.$$

Because $a<c$, it is enough to show that the nominator is positive. The nominator takes on the form $a(c-w)+(c^2-3cw+2z)$. Because $c > \frac{aw-z}{a-w} > \frac{z}{w} > w$, one has to show that $c^2-3cw+2z>0$. Determinant $\Delta = 9w^2-8z = 8w(1-q)+w^2 > 0$. So

$$\frac{3w+\sqrt{9w^2-8z}}{2} < \frac{z}{w} \Rightarrow \sqrt{9w^2-8z} < \frac{2z-3w^2}{w} \Rightarrow z(z-w^2) > 0.$$

Proof of ∗: we have

$$\frac{\partial}{\partial c} prob(a,c) = \frac{4(a^2 + a(c-3w) - cw + 2z)}{(a-c)^3} = 4\frac{c(a-w)+a^2-3aw+2z}{(a-c)^3}.$$

Because $a<c$ hence the denominator is negative, and the sign of the derivative depends on the sign of the nominator. The nominator is a linear function of c. It is enough to look at the sign of the nominator at the points $c = \frac{aw-z}{a-w}$ and $c = n$

For $c = \frac{aw-z}{a-w}$ the nominator equals

$$a^2 + a(n-3w) + 2z - nw.$$

This a quadratic function of a with zeros at

$$a_{1,2} = \frac{(3w-n) \pm \sqrt{(n-w)^2 + 8(w^2-z)}}{2}.$$

We show that

$$\frac{nw-z}{n-w} - 1 < a_2 < \frac{nw-z}{n-w}.$$

We have

$$\frac{nw-z}{n-w}-a_2=\frac{nw-z}{n-w}-\frac{(3w-n)+\sqrt{(n-w)^2+8(w^2-z)}}{2}$$
$$=\frac{n-w}{2}\left[1+2\frac{w^2-z}{(n-w)^2}-\sqrt{1+8\frac{w^2-z}{(n-w)^2}}\right]$$
$$=\frac{n(1-q)}{2}\left[1-2\frac{q}{n(1-q)}-\sqrt{1-8\frac{q}{n(1-q)}}\right].$$
$$(ĺ\,)$$

It is easy to see that $0<(ĺ\,)<1$ for $n>\left[\frac{q(1+q)}{1-q}\right]^2$, i.e. for $n\geq 2$.

Hence, for $c=n$ the nominator is negative, so the derivative is positive. The function $prob(a,c)$ attains its maximum at $c=\frac{aw-z}{a-w}$ or $c=n$. We show that for $a=0,1,\ldots,\left\lfloor\frac{aw-z}{a-w}\right\rfloor-1$ and $q>\beta^{-1}\left((a+1)+1,n-(a+1);\frac{1+\gamma}{2}\right)$ holds $prob\left(a,\frac{aw-z}{a-w}\right)>prob(a,n)$. We have

$$prob\left(a,\frac{aw-z}{a-w}\right)-prob(a,n)=\frac{4n(1-q)((n-1)q-a)}{(a-n)^2}.$$

For ``reasonable'' γ we have $\beta^{-1}\left((a+1)+1,n-(a+1);\frac{1+\gamma}{2}\right)>\frac{a}{n-1}$. Hence

$$prob\left(a,\frac{aw-z}{a-w}\right)>prob(a,n).$$

Numbers a and c are natural, so we obtain thesis from $\bullet$ and $*$.

Proof of the Theorem 2. For all $q\in(q_1,q_2)$ bounds $\lfloor x_1(q)\rfloor$ and $\lfloor x_2(q)\rfloor+1$ of confidence interval remains the same. Hence, the probability $prob(a,c)$ depends only on q. Because $w=nq$, $z=nq(1-q)+(nq)^2$, so

$$prob(a,c) = \frac{(a+c-2w)^2 + 4(z-w^2)}{(c-a)^2} = \frac{(a+c-2nq)^2 + 4nq(1-q)}{(c-a)^2}$$
$$= \frac{(a+c)^2 - 4n(a+c-1)q + 4n(n-1)q^2}{(c-a)^2}.$$

This is a quadratic function of q and it achieves its maximum at q_1 or q_2.

REFERENCES

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.

ZIELIŃSKI R. (2006) Exact distribution of the natural ARPR estimator in small samples from infinite populations, Statistics In Transition, 7, 881-888.

ZIELIŃSKI W. (2009), A nonparametric confidence interval for At-Risk-of-Poverty-Rate, Statistics in Transition new series, 10, 437-444.