STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 437—444

A NONPARAMETRIC CONFIDENCE INTERVAL FOR AT-RISK-OF-POVERTY-RATE

Wojciech Zieliński[1]

ABSTRACT

In the European Commission Eurostat document Doc. IPSE/65/04/EN page 11, the ``*at-risk-of-poverty rate*'' (ARPR) is defined as a percent of population with income smaller than 60% of population median. Zieliński (2008) proposed a distribution-free confidence interval for ARPR. In the paper, an example of application of the constructed confidence interval is shown.

Key words: binomial distribution, confidence interval, ARPR.

1. Introduction

In the European Commission Eurostat document Doc. IPSE/65/04/EN page 11, the ``*at-risk-of-poverty rate*'' (ARPR) is defined as follows. Let $EQ\_INC_i$ denote the equivalised disposable income of the *i*-th person and let $weight_i$ denote the weight of person *i*. The ``*at-risk-of-poverty threshold*'' (ARPT) is calculated as 60% of calculated median value, i.e.

$$\text{ARPT} = \text{At risk of poverty threshold} = 60\%\text{EQ\_INC}_{\text{MEDIAN}},$$

where

$$EQ\_INC_{MEDIAN} = \begin{cases} \frac{1}{2}(EQ\_INC_j + EQ\_INC_{j+1}), & \text{if } \sum_{i=1}^{j} weight_i = \frac{W}{2} \\ EQ\_INC_{j+1}, & \text{if } \sum_{i=1}^{j} weight_i < \frac{W}{2} < \sum_{i=1}^{j+1} weight_i \end{cases},$$

and

$$W = \sum_{All\ persons} weight_i.$$

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

Then the ``*at-risk-of-poverty rate*" is calculated as the percentage of persons (over the total population) with an equivalised disposable income below the *at-risk-of-poverty threshold* (i.e. the equivalised disposable income of each person is compared with *at-risk-of-poverty threshold*). The cumulated weights of persons whose equivalised disposable income is below the *at-risk-of-poverty threshold*, is divided by the cumulated weights of the total population (i.e. sum of all the personal weights):

$$ARPR = \frac{\sum_{All\ persons\ with\ EQ\_INC<at-risk-of-poverty\ threshold} weight_i}{W} \times 100\%.$$

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 and *Med* denotes the sample median. The estimator $\widehat{ARPR}$ is defined:

$$\widehat{ARPR} = \frac{1}{n} \#\{X_i \le 0.6 \cdot Med\},$$

where #*S* denotes the cardinality of the set *S*.

The properties of the above estimator (bias, variance etc.) depends strongly on the distribution *F* of population income. Zieliński (2006) showed that the estimator is almost unbiased, i.e.

$$E_F \widehat{ARPR} \approx ARPR = F(0.6 \cdot Q(0.5))$$

for all continuous *F* (*Q(0.5)* stands for the median of the distribution *F*). He also calculated its variance.

However, the problem is in interval estimation. Zieliński (2007) proposed a nonparametric confidence interval for *ARPR*. It appeared that his confidence interval is too conservative, i.e. the true confidence level is significantly larger that the nominal one. In what follows, we construct a confidence interval for *ARPR,* the confidence level of which is near nominal one.

2. Confidence interval

Let *F* denotes the cdf of a distribution of population income. It is assumed that *F* is continuous. We are interested in estimation of the parameter

$$\theta = F(\alpha Q(q)),$$

for given $\alpha, q \in (0,1)$, where $Q(\cdot)$ denotes the quantile function $(Q(x) = F^{-1}(x))$. For $\alpha = 0.6$ and $q = 0.5$ parameter θ is *ARPR*. We are interested in constructing a confidence interval for θ.

Let $X_1, X_2, \ldots, X_n$ be a sample from F and let $X_{1:n} < \cdots < X_{n:n}$ be order statistics. As an estimator of θ we take

$$\hat{\theta} = \frac{1}{n} \#\{X_i \le \alpha \cdot X_{M:n}\},$$

where $M = \lfloor qn \rfloor + 1$ ($\lfloor a \rfloor$ is the greatest integer not greater than *a*). Here, $X_{M:n}$ is an estimator of *q*-quantile *Q(q)* of the *F* distribution.

Let ξ be the number of observations not greater than $\alpha \cdot X_{M:n}$:

$$\xi = \#\{X_i \le \alpha \cdot X_{M:n}\}.$$

The distribution of ξ is

$$P_F\{\xi = k\} = P_F\{\xi \ge k\} - P_F\{\xi \ge k+1\} = P_F\{X_{k:n} \le \alpha \cdot X_{M:n}\}$$
$$- P_F\{X_{k+1:n} \le \alpha \cdot X_{M:n}\},\ k = 0, \ldots, M-1.$$

where (David and Nagaraja 2003)

$$P_F\{X_{k:n} \le \alpha \cdot X_{M:n}\} = \frac{n!}{(k-1)!(M-k-1)!(n-M)!} \int_{-\infty}^{\infty} (1-F(v))^{n-M} \cdot$$

$$f(v) \int_{-\infty}^{\infty} F(u)^{k-1} [F(v) - F(u)]^{M-k-1} f(u)\,du\,dv$$

$$= \frac{n!}{(k-1)!(M-k-1)!(n-M)!} \int_0^1 (1-v)^{n-M} \int_0^1 u^{k-1} [v-u]^{M-k-1}\,du\,dv$$

$$= \int_0^1 B_{k,M-k}\left(\frac{F(\alpha Q(v))}{v}\right) b_{M,n-M+1}(v)\,dv$$

Here, $B_{a,b}(\cdot)$ and $b_{a,b}(\cdot)$ denotes cdf and pdf of beta distribution with parameters *(a,b)*, respectively. That distribution may be written in the form

$$P_F\{X_{k:n} \le \alpha \cdot X_{M:n}\} = B_{k,M-k}\left(\frac{F(\alpha Q(q))}{q}\right)$$

$$+ \int_0^1 \left[B_{k,M-k}\left(\frac{F(\alpha Q(v))}{v}\right) - B_{k,M-k}\left(\frac{F(\alpha Q(q))}{q}\right) \right] b_{M,n-M+1}(v)\,dv.$$

It is well known, that if S_n is a random variable distributed as binomial with parameters n and p, then

$$P_{n,p}\{S_n \le k\} = \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} = B_{n-k,k+1}(1-p).$$

Hence, the distribution of ξ is almost binomial with parameters *M-1* and $\dfrac{F(\alpha Q(q))}{q}$.

If F is power distribution with shape parameter b, i.e.

$$F(x) = x^b, \quad 0 < x < 1,$$

then

$$\frac{F(\alpha Q(v))}{v} \equiv \alpha^b,$$

and

$$\int_0^1 \left[B_{k,M-k}\left(\frac{F(\alpha Q(v))}{v} \right) - B_{k,M-k}\left(\frac{F(\alpha Q(q))}{q} \right) \right] b_{M,n-M+1}(v)dv = 0.$$

Hence, in case of power function distribution, ξ is binomially distributed with parameters *M-1* and α^b. For this distribution $\theta = \alpha^b q$.

Let $\gamma \in (0,1)$. Consider an interval (see Appendix)

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

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

$$P_F\left\{ \theta \in \left(qB^{-1}\left(\xi, M-\xi+1; \frac{1-\gamma}{2} \right); qB^{-1}\left(\xi+1, M-\xi; \frac{1+\gamma}{2} \right) \right) \right\} \ge \gamma.$$

Hence, *(*)* is a confidence interval for θ. The question is: what is the value of

$$P_F\left\{ \theta \in \left(qB^{-1}\left(\xi, M-\xi+1; \frac{1-\gamma}{2} \right); qB^{-1}\left(\xi+1, M-\xi; \frac{1+\gamma}{2} \right) \right) \right\} \qquad (**)$$

for distributions F other the power function one? In general, it is impossible to calculate *(**)* because it strongly depends on F. In tables 1-3 there are calculated probabilities *(**)* (denoted by $\hat{\gamma}$ and the mean length ($\hat{\Delta}$) of confidence interval *(*)* for:

Pareto distribution with pdf $b\left(\frac{1}{x+1}\right)^{b-1}$ for $x \in (0,\infty)$;

Gamma distribution with pdf $\frac{1}{\Gamma(b)} x^{b-1} e^{-x}$ for $x \in (0,\infty)$;

Lognormal distribution with pdf $\frac{1}{xb\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x}{b}\right)^2\right]$ for $x \in (0,\infty)$.

Calculations were made for α=0.6, q=0.5, n=20, 200, 2000, $M = \lfloor n/2 \rfloor + 1$ and γ=0.95.

3. Conclusions

The confidence interval for $\theta = F(\alpha \cdot Q(q))$ may be constructed as a confidence interval in binomial distribution based on the number of observations not greater than $\alpha \cdot X_{M:n}$. Its confidence level for power function distribution is exactly the same as the confidence level of confidence interval for binomial proportion. For Pareto, Gamma and Lognormal distributions confidence level does not differ much from the nominal one. It may be expected that for other continuous distributions the confidence level will behave similarly.

Appendix: confidence interval for binomial proportion

Let η be a binomial random variable with parameters n and unknown p. It is well known that

$$P_p\{\eta \le k\} = B_{n-k,k+1}(1-p) \quad \text{and} \quad P_p\{\xi \ge k\} = B_{k,n-k+1}(p).$$

Let $\delta \in (0,1)$ be a given number. Confidence interval for p at the confidence level δ is defined as

$$P_p\{p_L(\eta) \le p \le p_U(\eta)\} = \delta, \quad \text{for all } p \in (0,1).$$

For given n and k let $p_L(k)$ be the solution of

$$B_{n-k+1,k}(1-p_L(k)) = \frac{1+\delta}{2} \quad \text{or equivalently} \quad B_{k,n-k+1}(p_L(k)) = \frac{1-\delta}{2}.$$

We obtain $p_L(k) = B^{-1}\left(k, n-k+1; \frac{1-\delta}{2}\right)$.

Similarly, we obtain $p_U(k) = B^{-1}\left(k+1, n-k; \frac{1+\delta}{2}\right)$.

Hence, the confidence interval for p at the confidence level δ is of the form

$$P_p\left\{B^{-1}\left(\eta, n-\eta+1; \frac{1-\delta}{2}\right) \le p \le B^{-1}\left(\eta+1, n-\eta; \frac{1+\delta}{2}\right)\right\} \ge \delta, \quad \text{for all } p \in (0,1).$$

The actual confidence level is higher than the nominal one because of discreteness of binomial distribution (see for example Brown et al. 2001).

Table 1. Pareto distribution

b	ARPR	n=20		n=200		n=2000	
		$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$
0.100	0.4738	0.9815	0.1804	0.9567	0.0473	0.9495	0.0136
0.111	0.4709	0.9761	0.1820	0.9476	0.0495	0.9448	0.0142
0.125	0.4672	0.9685	0.1835	0.9474	0.0521	0.9373	0.0148
0.143	0.4625	0.9917	0.1948	0.9259	0.0540	0.9433	0.0158
0.167	0.4565	0.9872	0.1991	0.9262	0.0580	0.9387	0.0168
0.200	0.4485	0.9795	0.2039	0.9093	0.0612	0.9311	0.0180
0.250	0.4377	0.9925	0.2170	0.9701	0.0656	0.9342	0.0196
0.333	0.4228	0.9849	0.2263	0.9442	0.0704	0.9358	0.0214
0.500	0.4024	0.9928	0.2435	0.9550	0.0776	0.9368	0.0234
1	0.3750	0.9800	0.2572	0.9531	0.0842	0.9453	0.0258
2	0.3585	0.9926	0.2705	0.9486	0.0870	0.9482	0.0269
3	0.3526	0.9908	0.2730	0.9578	0.0889	0.9465	0.0272
4	0.3496	0.9897	0.2742	0.9584	0.0893	0.9458	0.0273
5	0.3477	0.9890	0.2748	0.9543	0.0891	0.9494	0.0275
6	0.3465	0.9885	0.2753	0.9565	0.0896	0.9490	0.0276
7	0.3456	0.9576	0.2709	0.9573	0.0898	0.9527	0.0277
8	0.3450	0.9579	0.2712	0.9574	0.0899	0.9486	0.0276
9	0.3444	0.9582	0.2715	0.9574	0.0900	0.9485	0.0276
10	0.3440	0.9584	0.2717	0.9572	0.0900	0.9524	0.0278

Table 2. Gamma distribution

b	ARPR	n=20		n=200		n=2000	
		$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$
0.100	0.4751	0.9887	0.1818	0.9828	0.0462	0.9581	0.0134
0.111	0.4724	0.9851	0.1837	0.9544	0.0470	0.9627	0.0141
0.125	0.4691	0.9799	0.1858	0.9666	0.0497	0.9587	0.0148
0.143	0.4649	0.9965	0.1951	0.9714	0.0525	0.9527	0.0155
0.167	0.4595	0.9941	0.2000	0.9749	0.0562	0.9513	0.0166
0.200	0.4521	0.9894	0.2058	0.9750	0.0602	0.9533	0.0178
0.250	0.4416	0.9791	0.2124	0.9590	0.0642	0.9566	0.0195
0.333	0.4257	0.9911	0.2294	0.9665	0.0710	0.9544	0.0215
0.500	0.3986	0.9937	0.2498	0.9537	0.0786	0.9544	0.0242
1	0.3402	0.9600	0.2734	0.9563	0.0906	0.9514	0.0279
2	0.2668	0.9692	0.2916	0.9492	0.0959	0.9477	0.0297
3	0.2178	0.9649	0.2890	0.9484	0.0953	0.9433	0.0294
4	0.1813	0.9781	0.2866	0.9545	0.0932	0.9416	0.0285
5	0.1527	0.9802	0.2764	0.9518	0.0893	0.9432	0.0274
6	0.1297	0.9627	0.2609	0.9502	0.0851	0.9420	0.0260
7	0.1109	0.9807	0.2562	0.9623	0.0819	0.9461	0.0248
8	0.0952	0.9900	0.2493	0.9491	0.0767	0.9458	0.0235
9	0.0820	0.9748	0.2352	0.9473	0.0726	0.9452	0.0221
10	0.0709	0.9850	0.2297	0.9467	0.0685	0.9439	0.0209

Table 3. Lognormal distribution

b	ARPR	n=20		n=200		n=2000	
		$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$
0.100	$1.63 \; 10^{-7}$	1.0000	0.1542	1.0000	0.0181	0.9997	0.0018
0.111	$2.14 \; 10^{-6}$	0.9999	0.1542	0.9995	0.0181	0.9957	0.0018
0.125	$2.19 \; 10^{-5}$	0.9992	0.1541	0.9953	0.0180	0.9990	0.0019
0.143	0.0002	0.9942	0.1534	0.9993	0.0184	0.9944	0.0021
0.167	0.0011	0.9687	0.1494	0.9773	0.0193	0.9752	0.0033
0.200	0.0053	0.9895	0.1606	0.9728	0.0256	0.9534	0.0065
0.250	0.0205	0.9839	0.1797	0.9532	0.0415	0.9536	0.0122
0.333	0.0627	0.9866	0.2243	0.9513	0.0654	0.9382	0.0197
0.500	0.1535	0.9727	0.2743	0.9413	0.0884	0.9313	0.0271
1	0.3047	0.9748	0.2853	0.9435	0.0933	0.9390	0.0288
2	0.3992	0.9924	0.2472	0.9450	0.0777	0.9460	0.0240
3	0.4324	0.9931	0.2229	0.9682	0.0685	0.9518	0.0206
4	0.4492	0.9861	0.2060	0.9503	0.0600	0.9541	0.0183
5	0.4593	0.9936	0.1985	0.9720	0.0560	0.9530	0.0166
6	0.4661	0.9966	0.1926	0.9728	0.0518	0.9523	0.0153
7	0.4709	0.9829	0.1835	0.9663	0.0486	0.9548	0.0143
8	0.4745	0.9880	0.1810	0.9804	0.0463	0.9535	0.0135
9	0.4774	0.9913	0.1789	0.9736	0.0439	0.9577	0.0128
10	0.4796	0.9935	0.1770	0.9617	0.0415	0.9530	0.0121

REFERENCES

BROWN L. D, CAI T. T, DASGUPTA A. (2001) Interval estimation for Binomial proportion, Statistical Science, 16, 101–133.

DAVID H. A., NAGARAJA H. N. (2003) Order Statistics, Third Edition, Wiley.

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

ZIELIŃSKI R. (2007) A confidence interval for ARPR – "at-risk-of-poverty-rate", Statistics In Transition, 8, 217–222.