A nonparametric confidence interval for *At-Risk-of-Poverty-Rate*: an example of application

Wojciech Zieliński
Department of Econometrics and Statistics
Warsaw University of Life Sciences
Nowoursynowska 159, PL-02-776-Warszawa
wojtek.zielinski@statystyka.info

Abstract

In the European Commision 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

Keywords: binomial distribution, confidence interval, $ARPR$.
JEL: C14, C13

1. Introduction

In the European Commision 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.

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

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

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

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):

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

Let $X_1, \ldots, X_n$ be a sample of disposable incomes of randomly drawn n persons and Med denotes the sample median. The natural estimator $\widehat{ARPR}$ is defined:

$$\widehat{ARPR} = \frac{1}{n}\#\{X_i \leq 0.6 \cdot Med\}.$$

The properties of $\widehat{ARPR}$ were investigated by Zieliński (2006, 2007).

However, the problem is in interval estimation of $ARPR$. Zieliński (2008) proposed a nonparametric confidence interval for $ARPR$. This confidence interval is presented in Chapter 2. In Chapter 3 an example of application of the confidence interval is shown.

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

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

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 $X_{1:n} \leq \cdots \leq X_{n:n}$ be order statistics. As an estimator of θ we take

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

where $M = \lfloor qn \rfloor + 1$ ($\lfloor a \rfloor$ is the greatest integer not grater 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 grater than $\alpha \cdot X_{M:n}$:

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

The distribution of ξ is

$$\begin{aligned} P_F\{\xi = k\} &= P_F\{\xi \geq k\} - P_F\{\xi \geq k+1\} \\ &= P_F\{X_{k:n} \leq \alpha \cdot X_{M:n}\} - P_F\{X_{k+1:n} \leq \alpha \cdot X_{M:n}\},\ k = 0, \ldots, M-1. \end{aligned}$$

I may be checked that (David and Nagaraja 2003, Zieliński 2008)

$$\begin{aligned} &P_F\{X_{k:n} \leq \alpha \cdot X_{M:n}\} = \int_0^1 B_{k,M-k}\left(\frac{F(\alpha Q(v))}{v}\right) b_{M,n-M+1}(v)dv \\ &= B_{k,M-k}\left(\frac{F(\alpha Q(q))}{q}\right) + \int_0^1 \left[B_{k,M-k}\left(\frac{F(\alpha Q(v))}{v}\right) - B_{k,M-k}\left(\frac{F(\alpha Q(q))}{q}\right)\right] b_{M,n-M+1}(v)dv. \end{aligned}$$

Here $B_{a,b}(\cdot)$ and $b_{a,b}(\cdot)$ denotes cdf and pdf of beta distribution with parameters (a,b), respectively.

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

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

We obtain:

$$\begin{aligned} P_F\{\xi = k\} =& \binom{M-1}{k} p(q)^k (1-p(q))^{M-k-1} \\ &+ \int_0^1 \left[\binom{M-1}{k} p(v)^k (1-p(v))^{M-k-1} - \binom{M-1}{k} p(q)^k (1-p(q))^{M-k-1} \right] b_{M,n-M+1}(v)dv. \end{aligned}$$

where

$$p(v) = \frac{F(\alpha Q(v))}{v}.$$

Hence, the distribution of ξ is almost binomial with parameters $M-1$ and $\frac{F(\alpha Q(q))}{q}$.
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).
This interval may be considered as a confidence interval for θ (Zieliński 2008). It appears that the probability

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

of covering the true value of θ strongly depends on the underlying distribution.

3. An example

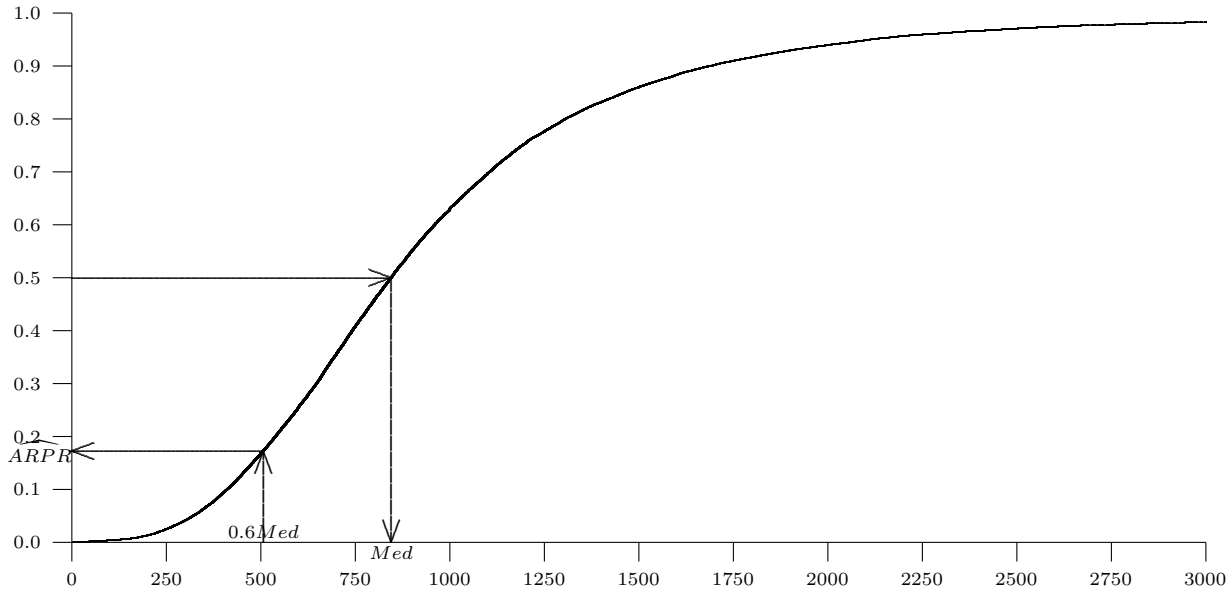
Theoretical results from Chapter 2 were applied to estimation of $ARPR$ in Poland in 2003. As a sample, there were 32292 data of the equivalised disposable income (data were used with kind permission of Dr Hanna Dudek, Warsaw University of Life Sciences). The empirical cdf of the most interesting part of the data are shown in the Picture.
It was calculated ($\alpha = 0.6, q = 0.5$):

$$M = 16148,\ Med = Q(q) = 845.0096904,\ \xi = 5576,\ \widehat{ARPR} = 0.17267.$$

Those calculations are illustrated in the Picture.
The confidence interval $(*)$ for $ARPR$ takes on the form ($\gamma = 0.95$)

$$\left(0.5B^{-1}(5576, 10573; 0.025); 0.5B^{-1}(5577, 10572; 0.975)\right) = (0.16903; 0.17633).$$

The question is: what is the confidence level of the above confidence interval? There are at least two methods of estimating that level. The first one relies on the fitting a theoretical distribution to given data. This



Picture. Empirical cdf

method is rather useless in the case. It is because, there are many different distributions F which may model considered data and for each such distribution calculations are numerically complicated and time consuming. The second method makes use of the well-known bootstrap technique (Chernick 1999). This method is simply and gives almost true results. So bootstrap method was applied in estimation of the confidence level of obtained confidence interval.

All data were numbered from 1 up to 32292. There were generated, according to uniform distribution on the set $\{1,\ldots,32292\}$, n numbers $i_1,\ldots,i_n$. From the of data, a sample $X_{i_1},\ldots,X_{i_n}$ was drawn. For the sample there were calculated the value of ξ. The procedure was repeated K times, so values $\xi_1,\ldots,\xi_K$ were obtained.

In the next step, mean value $\bar{\xi} = \frac{1}{Kn}\sum_{i=1}^{K}\xi_i$ was calculated. For every ξ_i, $i = 1,\ldots,K$, the confidence interval $(*)$ was calculated, and it was checked whether $\bar{\xi}$ falls into the obtained confidence interval or not. The percentage of confidence intervals containing $\bar{\xi}$ may be considered as an estimator of confidence level of confidence interval for $ARPR$.

In our investigations $n = 1000$ and $K = 500$. Obtained results are as follows.

$$\bar{\xi} = 0.17315,\ \sqrt{\frac{1}{K}\sum_{i=1}^{K}(\xi_i - \bar{\xi})^2} = 0.01064.$$

Estimated confidence level is 0.954 and the standard error of that estimate is 0.0094. Hence, it may be said that the obtained confidence interval for $ARPR$ in Poland in 2003 is on the confidence level about 0.954.

References

Brown L. D., Cai T. T., DasGupta A. (2001) Interval Estimation for Binomial proportion, Statistical Science, 16, 101-133

Chernick M. R. (1999), Bootstrap Methods, A Practitioner's Guide, Wiley

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

Zieliński W. (2008), A nonparametric confidence interval for At-Risk-of-Poverty-Rate, The Economics Letters (submitted) (preprint: http://wojtek.zielinski.statystyka.info/Inne_informacje/prace_pedefy/S46.pdf)

Appendix: confidence interval for binomial proportion

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

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

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

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

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

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

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