Wydział Nauk Ekonomicznych i Zarządzania
Uniwersytetu Szczecińskiego

Folia Oeconomica Stetinensia
DOI: 10.2478/v10031-008-0018-7

A REMARK ON THE INTERPRETATION OF POOLING RESULTS

Wojciech Zieliński, Prof.

Department of Econometrics and Statistics
Warsaw University of Life Sciences
Nowoursynowska 159,
02-787 Warszawa
e-mail:wojtek.zielinski@statystyka.info

Received 13 November 2008, Accepted 29 January 2009

Abstract

A multinomial distribution $(-1,\pi_1)$, $(0,\pi_2)$, $(1,\pi_3)$, $\pi_1+\pi_2+\pi_3=1$ is observed. Suppose that n_1 values -1, n_2 values 0 and n_3 values 1 ($n=n_1+n_2+n_3$) were observed. The construction of the confidence region for the vector (π_1,π_2,π_3) is given. Theoretical considerations are illustrated by an example of the results of a pool.

Keywords: multinomial distribution, confidence regions.

JEL classification: C13, C16.

Introduction

In Poland there are several pollsters. In the Table 1 there are given results of a Gallup Poll made by IPSOS in September 16, 2005.

Table 1. Results of a Gallup Poll made by IPSOS in September 16, 2005

Preferences in presidential election (%)	
Donald Tusk	43
Lech Kaczyński	23
Włodzimierz Cimoszewicz	17
Andrzej Lepper	7
Marek Borowski	3
Jarosław Kalinowski	2
Maciej Giertych	2
Janusz Korwin-Mikke	1
Henryka Bochniarz	1
Stanisław Tymiński	1

Source: www.ipsos.pl/3_5_070.html.

The pollster claims that "*the number of adults in a sample is* 959; *for an analogous random sample the statistical error of estimates is not greater than* (±)3.2% *at the confidence level* 0.95". Here appears a problem with the population interpretation of results. For example, if one wants to estimate the real support for the last of the candidates, then according to the comment the interval (-2.2%, 4.2%) is obtained. It suggests that this candidate may have a negative support (?!). How then should the results of a pool be generalized to the population?

In what follows we will consider a pool with three possible answers. Let X denote the random variable representing those answers. It may be assumed that X is multinomially distributed:

$$P_\theta = (X = -1) = \pi_1,$$

$$P_\theta = (X = 0) = \pi_2,$$

$$P_\theta = (X = 1) = \pi_3,$$

where

$$\theta = (\pi_1, \pi_2, \pi_3),$$

$$0 \le \pi_1, \pi_2, \pi_3 \le 1,$$

$$\pi_1 + \pi_2 + \pi_3 = 1.$$

Let us assume that in a sample of size n the value -1 was observed n_1 times, the value 0 - n_2 times and the value 1 - n_3 times. Of course $n = n_1 + n_2 + n_3$. It is known that the

maximum likelihood estimator of θ is ($p_1 = n_1 / n$, $p_2 = n_2 / n$, $p_3 = n_3 / n$). The problem is with the interval estimation of θ.

1. Confidence Region

There are a lot of papers devoted to the problem of simultaneous confidence intervals for the probabilities of multinomial distribution. An extensive review of construction methods may be found in Biszof[1], Correa[2], May[3]. General rule of construction is based on the set of inequalities

$$\frac{|p_i - \pi_i|}{\pi_i(1-\pi_i)} \le c, \quad i = 1,2,3 ,$$

where c is a constant such that

$$P_\theta\{\frac{|p_i - \pi_i|}{\pi_i(1-\pi_i)} \le c, \quad i = 1,2,3\} = 1-\alpha, \quad \forall\theta.$$

Those confidence regions are easy to calculate. However, simultaneous confidence intervals have at least two disadvantages. Firstly, the obtained confidence intervals may go out of the (0, 1) interval and secondly, in their construction the condition $\pi_1 + \pi_2 + \pi_3 = 1$ was not exploited.

For example, let the following sample be given: $n_1 = 1$, $n_2 = 1$, $n_3 = 48$. In the Table 2 there are given limits of some of the known simultaneous confidence intervals ($1-\alpha = 0.95$).

Table 2. Limits of some of the known simultaneous confidence intervals ($1-\alpha = 0.95$)

	QH		GM		NB		FS	
p_1=0.02	0.0025	0.1402	0.0026	0.1361	-0.1493	0.1893	-0.1303	0.1703
p_2=0.02	0.0025	0.1402	0.0026	0.1361	-0.1493	0.1893	-0.1303	0.1703
p_3=0.96	0.8300	0.9916	0.8340	0.9914	0.7907	1.1293	0.8097	1.1103

Source: own calculations.

Here QH denotes Quesenberry[4] construction:

$$n_i(p_i - \pi_i)^2 \le \chi^2(\alpha;2)\pi_i(1-\pi_i), \quad i = 1,2,3 ,$$

GM denotes Goodman[5] construction:

$$n_i(p_i-\pi_i)^2 \le \chi^2(\alpha/3;1)\pi_i(1-\pi_i), \quad i=1,2,3,$$

NB denotes *naive binomial* construction:

$$n_i(p_i-\pi_i)^2 \le \chi^2(\alpha;1)\left(\frac{1}{4}\right), \quad i=1,2,3,$$

FS denotes Fitzpatrick[6] construction:

$$n_i(p_i-\pi_i)^2 \le \gamma, \quad i=1,2,3,$$

where $\gamma=1$ for $\alpha=0.1$, $\gamma=1.13$ for $\alpha=0.05$ and $\gamma=1.40$ for $\alpha=0.01$.

Note that the left ends of some of the calculated confidence intervals are negative or the sum of admissible probabilities is greater than one.

Another approach to the problem of the construction of confidence region is the application of Pearson statistic connected with classical chi-square test. This statistic is given by a formula:

$$n\left(\frac{(p_1-\pi_1)^2}{\pi_1}+\frac{(p_2-\pi_2)^2}{\pi_2}+\frac{(p_3-\pi_3)^2}{\pi_3}\right).$$

Applying the conditions $\pi_1+\pi_2+\pi_3=1$ and $p_1+p_2+p_3=1$ we obtain:

$$n\left(\frac{(p_1-\pi_1)^2}{\pi_1}+\frac{(p_2-\pi_2)^2}{\pi_2}+\frac{((p_1+p_2)-(\pi_1+\pi_2))^2}{1-\pi_1-\pi_2}\right).$$

A confidence region is given by a solution with respect to (π_1, π_2) of an inequality:

$$n\left(\frac{(p_1-\pi_1)^2}{\pi_1}+\frac{(p_2-\pi_2)^2}{\pi_2}+\frac{((p_1+p_2)-(\pi_1+\pi_2))^2}{1-\pi_1-\pi_2}\right)<\chi,$$

where $\chi=\chi^2(\alpha;2)/n$.

Let

$$p_2^L(\pi_1)=-\frac{\chi(-1+\pi_1)\pi_1+\pi_1^2+p_1^2-2\pi_1(p_1+p_2-p_1p_2)-\sqrt{\Delta(\pi_1)}}{2(\pi_1+\chi\pi_1-p_1^2)},$$

$$p_2^P(\pi_1)=-\frac{\chi(-1+\pi_1)\pi_1+\pi_1^2+p_1^2-2\pi_1(p_1+p_2-p_1p_2)+\sqrt{\Delta(\pi_1)}}{2(\pi_1+\chi\pi_1-p_1^2)},$$

where

$$\Delta(\pi_1) = \left(\chi(-1+\pi_1)\pi_1 + \pi_1^2 + p_1^2 + 2\pi_1(p_1(-1+p_2) - p_2)\right)^2 + 4(-1+\pi_1)\pi_1(\pi_1 + \chi\pi_1 - p_1^2)p_2^2.$$

Let p_1^L and p_1^P be such numbers that $\Delta(p_1^L) = 0$ and $\Delta(p_1^P) = 0$:

$$p_1^L = \frac{\chi + 2p_1 - \sqrt{\chi}\sqrt{\chi + 4p_1 - 4p_1^2}}{2(1+\chi)}; \qquad p_1^P = \frac{\chi + 2p_1 + \sqrt{\chi}\sqrt{\chi + 4p_1 - 4p_1^2}}{2(1+\chi)}.$$

The confidence region is given by:

$$\left\{(\pi_1, \pi_2) \in (0,1)\times(0,1) : p_1^L \le \pi_1 \le p_1^P,\ p_2^L(\pi_1) \le \pi_2 \le p_2^P(\pi_1)\right\}.$$

The value of π_3 is calculated from the equality $\pi_1 + \pi_2 + \pi_3 = 1$.

2. Example

Let us consider a modified version of the results of the poll presented in the Introduction given in Table 3.

Table 3. A modified version of the results of the poll presented in the Introduction

Preferences in presidential election	
Donald Tusk	43%
Lech Kaczyński	23%
Other candidate	34%

Source: own calculations.

In the Figure 1 there is shown the confidence region for the probability π_1 (x-axis) and the π_2 (y-axis) of support for the first and the second candidate respectively. The third probability π_3 is calculated as $\pi_3 = 1 - \pi_1 - \pi_2$. The values of the point estimators of π_1 and π_2 are marked with a dot.

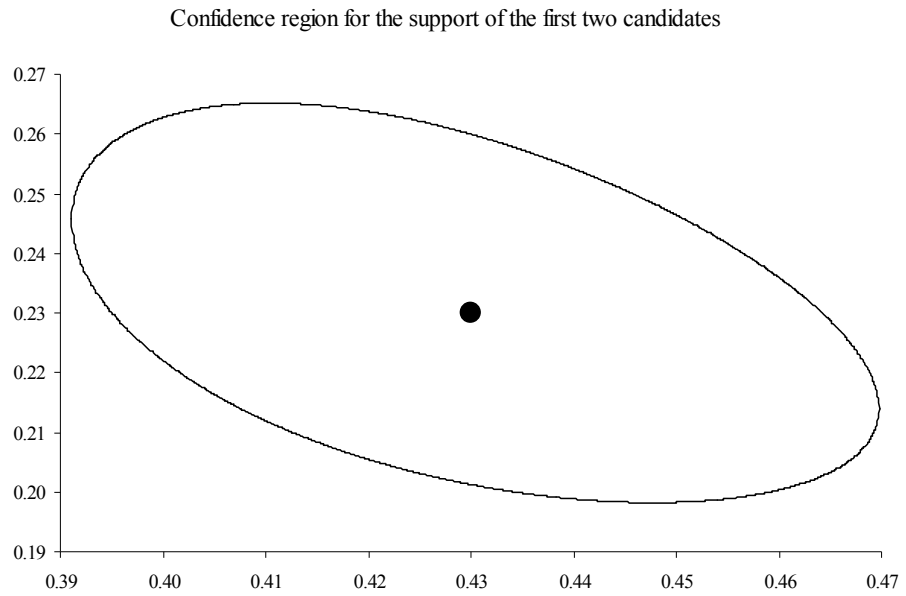

Fig. 1. Confidence region for the support of the first two candidates
Source: own study.

Notes

[1] Biszof, Mejza (2004).

[2] Correa, (2001).

[3] May, Johnson (1997).

[4] Quesenberry, Hurst (1964).

[5] Goodman (1965).

[6] Fitzpatrick, Scott (1987).

References

Biszof, A. & Mejza, S. (2004). Jednoczesne przedziały ufności dla prawdopodobieństwa w rozkładzie wielomianowym. *Colloquium Biometryczne.* 34, 77-84.

Correa, J. C. (2001). Interval Estimation of the Parameters of the Multinomial Distribution, http://ip.statjournals.net:2002/InterStat/ARTICLES/2001/articles/O01001.pdf.

Fitzpatrick, S. & Scott, A. (1987). Quick Simultaneous Confidence Intervals for Multinomial Proportions. *Journal of the American Statistical Association*. 82, 875-878.

Goodman, L. A. (1965). On Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics.* 7, 247-254.

May, W. L. & Johnson, W. D. (1997). Properties of Simultaneous Confidence Intervals for Multinomial Proportions. *Communications in Statistics – Simulations.* 26, 495-518.

Quesenberry, C. P. & Hurst, D.C. (1964). Large Sample Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics.* 6, 191-195.