STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 535-544

STATISTICAL ANALYSIS OF A QUESTIONNAIRE: VOLUNTARY HEALTH INSURANCE IMPLEMENTATION AMONG PATIENTS SUFFERING FROM ALLERGY AND ASTHMA

Marta Zalewska[1]
Wojciech Zieliński[2]

ABSTRACT

We consider statistical analysis of multiple answers in a questionnaire. We propose a new method of calculating simultaneous confidence regions. In a communication presented at the European Academy of Allergy and Clinical Immunology the authors (Borowicz et al. (2009)) reported the proportions of respondents which gave one of three possible exclusive answers in a questionnaire concerning the role of voluntary health insurance. There were three possible answers. Apart from percentages of answers confidence intervals of every single answer have been reported. Unfortunately inference about the population based on such intervals may lead to imprecise conclusions.

The inference about the respective population suffering from allergy and asthma proportions requires the construction of two-dimensional confidence region. We propose the use of a simultaneous confidence intervals to inference about true population proportions.

Most of our attention is given to the case of three possible answers but the results may be generalized to any questionnaire with more than two excluding answers.

Key words: confidence region, health insurance, multiple responses, questionnaire

[1] Department of the Prevention of Environmental Hazards and Allergology, Medical University of Warsaw, Banacha 1a, 02-097 Warszawa,e-mail: zalewska.marta@gmail.com

[2] Department of Econometrics and Statistics, Warsaw University of Life Sciences, Nowoursynowska 159, 02-776 Warszawa, e-mail: wojciech_zielinski@sggw.pl

1. Introduction

We consider statistical analysis of multiple answers in a questionnaire. We propose a new method of calculating simultaneous confidence regions. In our article we concentrate on the example of voluntary health insurance implementation among patients suffering from allergy and asthma. There is an obligatory health insurance system in Poland. Unfortunately this system does not work efficiently, mainly because of incorrect diversification of funds. For these reasons together with the obligatory health insurance system we have the optional voluntary health insurance system (VHI) based on voluntary premium. Quite a large number of people in Poland participate in VHI system but the reasons for participating in this system are different. Epidemiology of Allergic Disease survey in Poland (presented during European Academy of Allergy and Clinical Immunology congress in 2009) included a question about the reasons for participating in VHI with three possible answers: additional, supplementary and substitutive (question number 566, and answers 566_1, 566_2 and 566_3 respectively). The results of the questionnaire given in Borowicz et al. (2009) are presented in Table 1.

Table 1. Results of the questionnaire (Borowicz et al. (2009))

The role of voluntary health insurance (question 566)	Frequency	Percentage
additional-increasing health service standard (answer 566_1)	1653	36.5
supplementary-expanding range of health service (answer 566_2)	1668	36.9
substitutive-enabling abandonment of public health care (answer 566_3)	1205	26.6

The results were obtained on the basis of the questionnaire based on the International Study of Asthma and Allergies in Childhood and the European Community Respiratory Health Survey II ECRHS II. All investigated subjects were randomly selected from PESEL (Personal Identification Number). Data acquisition was done by the Computer Assisted Personal Interviewing with GSM transmission to update the main database at the Medical University of Warsaw (http://ecap.pl/eng_www).

The question is: what are the population suffering from allergy and asthma percentages π_1, π_2 and π_3 of the Polish citizens participating in VHI system from the appropriate reasons (additional, supplement and substitutive). The standard approach is to construct individual confidence intervals. Unfortunately this approach may lead to wrong conclusions. Therefore, in what follows we propose to construct a confidence region for percentages π_1, π_2 and π_3 simultaneously.

2. Statistical model

Let X denote the random variable describing answers. It may be assumed that X is multinomially distributed:

$$P_{\boldsymbol{\pi}}\{X=1\}=\pi_1, P_{\boldsymbol{\pi}}\{X=2\}=\pi_2, P_{\boldsymbol{\pi}}\{X=3\}=\pi_3,$$

where $\boldsymbol{\pi}=(\pi_1,\pi_2,\pi_3)$ and $0<\pi_1,\pi_2,\pi_3<1$, $\pi_1+\pi_2+\pi_3=1$. Values of X symbolize answers to questions in the questionnaire (i.e. $X=1$ means that the answer is 566_1, $X=2$ means that the answer is 566_2 and $X=3$ - the answer 566_3). Probabilities π_1, π_2 and π_3 are (multiplied by 100%) population suffering from allergy and asthma (population to be short) percentages of obtaining answers to the questions.

Assume that in a sample of size n, value 1 was observed n_1 times, value 2 - n_2 times and value 3 - n_3 times. Of course $n_1+n_2+n_3=n$. It is known that the maximum likelihood estimator of $\boldsymbol{\pi}$ is: $\hat{\pi}_1=n_1/n$, $\hat{\pi}_2=n_2/n$ and $\hat{\pi}_3=n_3/n$. The problem is in the interval estimation of $\boldsymbol{\pi}$, the vector comprising the probabilities of answers.

In standard approach, each of the probabilities is estimated separately. It means, that three confidence intervals are obtained, usually on the basis of normal approximation, i.e. a confidence interval of the form is built for π_i (at the confidence level $1-\alpha$)

$$\left(\hat{\pi}_i-\frac{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}{\sqrt{n}}z, \hat{\pi}_i+\frac{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}{\sqrt{n}}z\right),$$

where z is the quantile of the order $1-\alpha/2$ of the standard normal distribution (i.e. $N(0,1)$ distribution) and

$$\hat{\pi}_i=\frac{n_i}{n}.$$

This approach gives the results (at 95% confidence level, i.e. $1-\alpha=0.95$) presented in Table 2.

Table 2. Individual confidence intervals for percentages

	Frequency n_i	Estimated Percentage $\hat{\pi}_i$	Left end	Right end
π_1 (answer 566_1)	1653	36.5	35.12	37.93
π_2 (answer 566_2)	1668	36.9	35.45	38.28
π_3 (answer 566_3)	1205	26.6	25.34	27.94

Classical inference is such that the population percentage of the answers to the first question is any number between 35.12% and 37.93%; to the second question - the number from the interval $(35.45\%, 38.28\%)$ and to the third one is from the interval $(25.34\%, 27.94\%)$. But this kind of inference may lead to wrong conclusions. Namely, it may be stated, that the percentage of population answers to the question 566_1 is 36%, to the second question (566_2) is also 36% and to the third question is 26% (i.e. $(\pi_1, \pi_2, \pi_3) = (0.36, 0.36, 0.26)$). Summing up those three values one obtains 98% of the population instead of expected 100% (2% of population is "missed"!). The other situation is also possible, i.e. stated population percentages may give more than 100% (for example: the percentage of answers to the first question is 37%, to the second - 38% and to the third question 27%). It appears also that the real confidence level of such conclusion is less than the nominal 95%. It means that the risk of wrong conclusions is too high: it is greater than the nominal 5%.

We are interested in simultaneous interval estimation of probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$.

3. Confidence region

There are a lot of papers devoted to the problem of simultaneous confidence intervals for probabilities of multinomial distribution. An extensive review of construction methods may be found in Biszof and Mejza (2004), Correa (2001), May and Johnson (1997). The general rule of construction is based on the set of inequalities

$$\frac{|\hat{\pi}_i - \pi_i|}{\sqrt{\pi_i(1-\pi_i)}} \leq c, \quad i = 1, 2, 3,$$

where c is a constant such that the following equality holds

$$P_{\boldsymbol{\pi}}\left\{\frac{|\hat{\pi}_i - \pi_i|}{\sqrt{\pi_i(1-\pi_i)}} \leq c, \quad i = 1, 2, 3\right\} = 1 - \alpha, \qquad \forall \boldsymbol{\pi}.$$

Those confidence regions are easy to calculate. However, simultaneous confidence intervals have two disadvantages. Firstly, the obtained confidence intervals may go out of $(0, 1)$ 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 Table 3 the limits of some of known simultaneous confidence intervals ($1 - \alpha = 0.95$) are given.

Table 3. Simultaneous confidence intervals

	QH		GM		NB		FS	
$\hat{\pi}_1$=0.02	0.0025	0.1402	0.0026	0.1361	-0.1493	0.1893	-0.1303	0.1703
$\hat{\pi}_2$=0.02	0.0025	0.1402	0.0026	0.1361	-0.1493	0.1893	-0.1303	0.1703
$\hat{\pi}_3$=0.96	0.8300	0.9916	0.8340	0.9914	0.7907	1.1293	0.8097	1.1103

QH denotes Quesenberry and Hurst (1964) construction:

$$n_i(\hat{\pi}_i - \pi_i)^2 \leq \chi^2(\alpha, 2)\pi_i(1 - \pi_i), \quad i = 1, 2, 3.$$

GM denotes Goodman (1965) construction:

$$n_i(\hat{\pi}_i - \pi_i)^2 \leq \chi^2(\alpha/3, 1)\pi_i(1 - \pi_i), \quad i = 1, 2, 3.$$

NB denotes naive binomial construction:

$$n_i(\hat{\pi}_i - \pi_i)^2 \leq \chi^2(\alpha, 1)(1/4), \quad i = 1, 2, 3.$$

FS denotes Fitzpatrick and Scott (1987) construction:

$$n_i(\hat{\pi}_i - \pi_i)^2 \leq \gamma, \quad i = 1, 2, 3.$$

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.

In what follows we propose another way of inference. We show how to built a confidence region for all three percentages simultaneously, such that:
1. all percentages in the confidence region will sum up to 100%;
2. the confidence level of conclusion will be equal to the nominal one.

Let us start with the very well known chi-square statistic Bland (2000), Greenwood and Nikulin (1996), Peacock and Peacock (2011) of the Pearson goodness-of-fit test:

$$\chi^2 = n \cdot \left(\frac{\left(\frac{n_1}{n} - \pi_1\right)^2}{\pi_1} + \frac{\left(\frac{n_2}{n} - \pi_2\right)^2}{\pi_2} + \frac{\left(\frac{n_3}{n} - \pi_3\right)^2}{\pi_3} \right).$$

To satisfy the first requirement the statistic above is transformed to

$$\chi^2(\pi_1, \pi_2) = n \cdot \left(\frac{\left(\frac{n_1}{n} - \pi_1\right)^2}{\pi_1} + \frac{\left(\frac{n_2}{n} - \pi_2\right)^2}{\pi_2} + \frac{\left(\frac{n_3}{n} - (1 - \pi_1 - \pi_2)\right)^2}{(1 - \pi_1 - \pi_2)} \right).$$

This statistic may be used in the construction of the confidence region in the following way. Let $\chi^2(\alpha; 2)$ denote the chi-square critical value with two degrees of freedom and the confidence level $1 - \alpha$. Then the confidence region for $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ is obtained as a solution with respect to $\boldsymbol{\pi}$ of the inequality $\chi^2(\pi_1, \pi_2) < \chi^2(\alpha; 2)$:

$$\left\{(\pi_1, \pi_2, \pi_3) : \chi^2(\pi_1, \pi_2) < \chi^2(\alpha; 2), \pi_3 = 1 - \pi_1 - \pi_2\right\}.$$

The theoretical background for constructing a confidence region for probabilities $\boldsymbol{\pi}$ may be found in Harton and Zieliński (2005) and Zieliński (2008). Explicit formulae for the confidence region may also be found in those papers. Note that the above construction tacitly assumes that the population in question is infinite. Of course, this is not exactly true because the population of adults in Poland is finite, but it is sufficiently large to accept this assumption as a reasonable approximation. Some remarks on the application of statistical methods devoted to the analysis of infinite populations to finite ones may be found in Zieliński (2011).

4. Results

We apply the constructed above confidence region to the problem of estimating the role of voluntary health insurance (question 566). In the questionnaire the $n = 4526$ answers were obtained. Among them there were $n_1 = 1653$ answers to the first question, $n_2 = 1668$ answers to the second question and $n_3 = 1205$ answers to the third one. As a confidence level 95% were taken, so the critical value of the chi-square distribution with two degrees of freedom equals 5.99. After some calculations the confidence region for π_1 and π_2 was obtained and is presented in Figure 1 (all computations were done using R-project with statistical computing (R Development Core Team (2008)); the computer codes in R were written by ourselves - see Appendix).

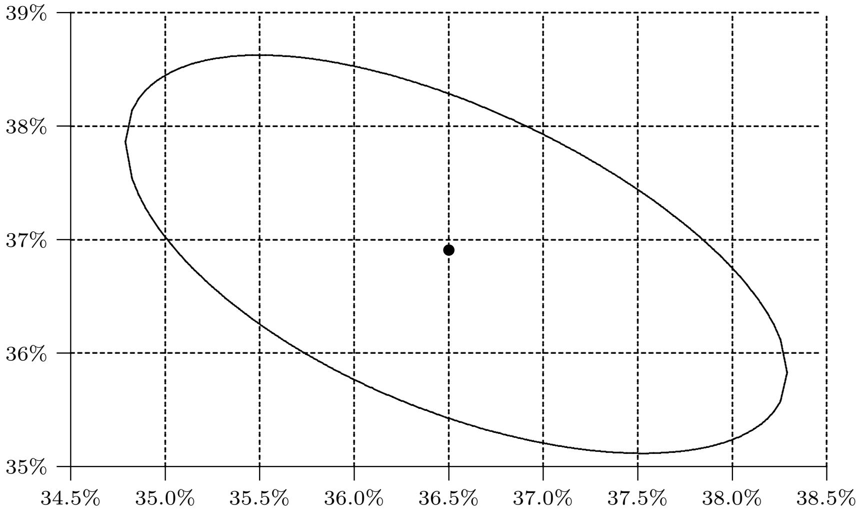

Figure 1. Confidence region for frequency of opinions of the role of health insurance

The letters π_1, π_2 and π_3 denote the proportion of answers to the first, second and third question, respectively in the population of interest (the graph shows only π_1 and π_2 because $\pi_3 = 1 - \pi_1 - \pi_2$). Recall that the first answer is “additional-increasing health service standard”; the second answer - “supplementary-expanding range of health service” and the third answer is “substitutive - enabling abandonment of public health care”.

The confidence region for π_1 and π_2 is inside the contour presented in Figure 1. We have to remember that this two dimensional graphs in fact inform us about three proportions (three possible answers to a given question). The dot “in the center of the confidence region” corresponds to the proportions in the sample: $\hat{\pi}_1 = 0.365$ (36.5%), $\hat{\pi}_2 = 0.369$ (36.9%)- these two are presented in the graph - and $\hat{\pi}_3$ given by $\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2 = 0.266$ (26.6%).

The interpretation of this confidence region is similar to that of a confidence interval but two-dimensional. Roughly speaking we trust that combinations of values π_1 and π_2 lie inside the region. More precisely, we make statements with probability of error 0.05 for all three proportions together. For example, a combination of proportions $\pi_1 = 0.375$ (37.5%) and $\pi_2 = 0.36$ (36%) may be true with high confidence because the point with coordinates 0.375 (37.5%) and 0.36 (36%) lies inside the contour shown in Figure 1 (i.e. $(\pi_1, \pi_2, \pi_3) = (0.36, 0.375, 0.265)$). On the other hand the combination $\pi_1 = 0.355$ and $\pi_2 = 0.36$ we treat as extremely unlikely because the point with coordinates 0.355 (35.5%) and 0.36 (36%) lies outside the contour. This last statement is true in spite of the fact that $\pi_1 = 0.355$ (35.5%) considered separately is possible and $\pi_2 = 0.36$ (36%) considered separately is also possible, but both these values together are unlikely (see Figure 2).

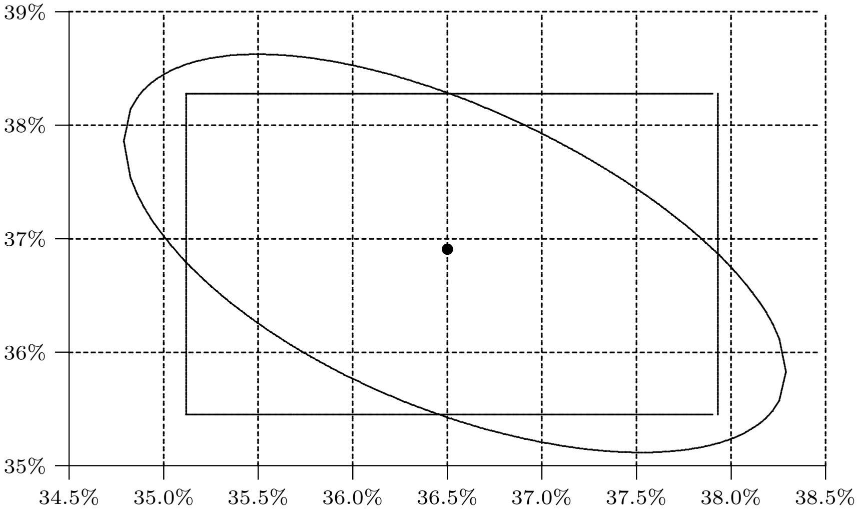

Figure 2. Confidence region for frequency of opinions of the role of health insurance. The rectangle shows two confidence intervals separately for π_1 and π_2.

In Figure 2 the elliptical confidence region is drawn along with the rectangular region of standard approximate confidence intervals. The analysis we present gives more precise information than conventional onedimensional confidence intervals. It might be argued that the interpretation of two-dimensional confidence regions is more difficult than that of one dimensional confidence intervals. However, easily accessible modern computer graphics allows us to show the relations between two or three variables and to understand and explain to the users the meaning of the confidence region.

5. Conclusions

The construction of the confidence region for three probabilities may be easily generalized to the problem of estimating more than three percentages. Of course, if there is a problem of estimating more than three proportions, the graphical illustration is impossible. For k possible mutually excluding answers it is sufficient to consider the statistic

$$\chi^2(\pi_1,\ldots,\pi_k) = n\cdot\sum_{i=1}^{k}\left(\frac{\left(\frac{n_i}{n}-\pi_i\right)^2}{\pi_i}\right)$$

and as the confidence region at the confidence level $1-\alpha$

$$\left\{(\pi_1,\ldots,\pi_k) : \chi^2(\pi_1,\ldots,\pi_k) < \chi^2(\alpha;k-1), \pi_1+\cdots+\pi_k = 1\right\}.$$

The interpretation is to some extent more complicated than in the case of individual confidence intervals, but it avoids the errors of inference.

In many allergological questionnaires there are numerous questions with multiple answers. We show that simultaneous inference is more appropriate and more informative than the one-dimensional ones. The latter can lose some relevant information while multidimensional analysis is more accurate.

REFERENCES

BISZOF, A., MEJZA, S., (2004). Jednoczesne przedziały ufności dla prawdopodobieństwa w rozkładzie wielomianowym, Colloquium Biometryczne, 34, 77-84.

BLAND, M., (2000). An introduction to medical statistics, Oxford University Press, Third edition.

BOROWICZ, J., SAMOLIŃSKI, B., FURMAŃCZYK, K., WALKIEWICZ, A., LUSAWA, A., MARSZAŁKOWSKA, J. et al., (2009). Attitudes towards the idea of voluntary health insurance implementation among patients suffering from allergy and asthma. Allergy 64 (Suppl. 90): 435 (abstract 1140)

CORREA, J. C., (2001). Interval Estimation of the Parameters of the Multinomial Distribution, 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.

GREENWOOD, D. E., NIKULIN, M. S., (1996). A guide to chi-squared testing, Willey.

HARTON, A., ZIELIŃSKI, W., (2005). Confidence region for probabilities of a multinomial distribution. Colloquium Biometricum, 35, 141-145.

MAY, W. L., JOHNSON, W. D., (1997). Properties of simultaneous confidence intervals for multinomial proportions, Communications in Statistics - Simulations, 26, 495-518.

PEACOCK, J. L., PEACOCK, P. J., (2011). Oxford handbook of medical statistics. Oxford University Press.

QUESENBERRY, C. P., HURST, D. C., (1964). Large sample simultaneous confidence intervals for multinomial proportions,Technometrics, 6, 191-195.

ZIELIŃSKI, W., (2008). A remark on interpretation of pooling results. Folia Oeconomica Stetinensia, 7(15), 56-62.

ZIELIŃSKI, W., (2011). Comparison of confidence intervals for fraction in finite populations, Quantitative Methods in Economics, XII, 177-182.

R DEVELOPMENT CORE TEAM, (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, http://www.R-project.org.

http://ecap.pl/eng_www/index_home.html.

Appendix. The computer code in R which was employed to draw Figure 1.

```
n1=1653 #input n1
n2=1668 #input n2
n3=1205 #input n2
n=n1+n2+n3 #the overall number of observations
alpha=0.05 #confidence level
pu1=as.vector(prop.test(n1,n,conf.level =1-alpha/2)$conf.int)
pu2=as.vector(prop.test(n2,n,conf.level =1-alpha/2)$conf.int)
prop.test(n3,n,conf.level =1-alpha/2)
p1=n1/n #estimated proportion $\pi_1$
p2=n2/n #estimated proportion $\pi_2$
p3=n3/n #estimated proportion $\pi_3$
chi=qchisq(1-alpha,2)/n #chi-square critical value
# assistant functions
delta= function(pi1){
delta=(chi*(-1+pi1)*pi1+pi1^2+p1^2+2*pi1*(p1*(-1+p2)-p2))^2+
4*(-1+pi1)*pi1*(pi1+chi*pi1-p1^2)*p2^2
delta
}
p2L=function(pi1){
p2L=-(chi*(-1+pi1)*pi1+pi1^2+p1^2-2*pi1*(p1+p2-p1*p2)+
sqrt(delta(pi1)))/ (2*(pi1+chi*pi1-p1^2))
p2L
}
p2P=function(pi1){
p2P=-(chi*(-1+pi1)*pi1+pi1^2+p1^2-2*pi1*(p1+p2-p1*p2)-
sqrt(delta(pi1)))/(2*(pi1+chi*pi1-p1^2))
p2P
}
P1L=(chi+2*p1-sqrt(chi)*sqrt(chi+4*p1-4*p1^2))/(2*(1+chi))
P1P=(chi+2*p1+sqrt(chi)*sqrt(chi+4*p1-4*p1^2))/(2*(1+chi))
c(P1L,P1P)
pi1=seq(P1L,P1P,length.out =250)
pi2L=p2L(pi1)
pi2P=p2P(pi1)
#Plot of confidence regions for frequencies ($\pi_1$,$\pi_2$)
ymax=max(pi2P)
ymin=min(pi2L)
plot(pi1,pi2L,xlim=c(P1L,P1P),ylim=c(round(ymin,2),round(ymax,2)),type=”l”,las=1,
ylab=expression(pi[2]),xlab=expression(pi[1]),lwd=2)
lines(pi1,pi2P,lwd=2)
points(p1,p2,pch=16)
```