ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Czesław Domański**

SOME REMARKS ON STATISTICAL INFERENCE FOR COMPLEX SAMPLES

Abstract

Classic theory of statistical inference gives us methods and verification of hypothesis for simple samples (observations are stochastically independent and have the same distribution). Because of costs and effectiveness of research we use simple samples. Observations in these samples are stochastically dependent and have different distribution.

The paper presents problems in estimation and verifications of hypothesis of consistency of distributions for complex samples.

Key words: complex samples, estimation, goodness of fit tests.

I. INTRODUCTION

Classic theory of statistical inference provides us estimation methods of unknown distribution parameters estimating the form of function which defines this distribution and hypotheses verification on the grounds of simple samples, that is such hypotheses in which observations are stochastically independent and have the same probability distribution. In general, however, we use complex samples with regard to costs and efficiency of research. Results of observations in these samples are realizations of stochastically dependent variates of various distributions. In representative research we distinguish among others the following schemes:

– dependent sampling (without replacement) with various choice probabilities,

– stratified sampling,

– cluster sampling,

– cluster and multistage sampling.

* Professor, Chair of Statistical Methods, University of Łódź.

For example, sampling without replacement eliminates stochastic independence of observation, stratification process causes diversification of choice probabilities of sample elements but multistage sampling influences the diversification of distribution.

This paper deals with problems connected with estimation, especially adaptation of methods of central limit theorem for complex samples and verification of goodness of fit for complex samples.

II. LIMIT THEOREMS

Representative method deals with procedures of sampling from finite populations and estimating on the grounds of obtained samples of unknown parameters in these populations. Since populations are finite, therefore samples must also be finite. What is more, if N – denotes the size of general population and n – sample size, then it is very reasonable to consider these situations in which $n<N$ (cases when $n=N$ are not the object of interest of sampling method). Economic and organizational considerations force statisticians to replace simple samples (simple sample means that each observation has the same distribution as the distribution of investigated variable in population) with complex samples. This fact makes using limit theorems for complex samples impossible.

In case of sampling without replacement the condition $n<N$ must be satisfied. That is why we can not use limit theorems known from probability calculus in which it is assumed that $n \to \infty$. We will mention here Lindeberg-Feller theorem, see Fisz (1967).

Theorem 1 (Lindeberg-Feller). Let $\{Y_k\}(k=1,2,...)$ be the sequence of independent variables. Let μ_k and $\tau_k>0$ denote an expected value and a standard deviation respectively and $G_k(y)$ – its distribution function.

$$C_n = \sqrt{\sum_{k=1}^{n} \sigma_k^2},$$

$$Z_n = \frac{1}{C}\sum_{k=1}^{n} (Y_k - \mu_k),$$

$$\Phi(z) = \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{z} e^{-\frac{\lambda^2}{2}} dt,$$

and $F_n(z)$ denotes distribution function of variate Z_n.

Necessary and sufficient condition to

$$\lim_{n\to\infty} \max_{1\leqslant k\leqslant aC_n} \frac{\sigma_k}{} = 0, \qquad \lim_{n\to\infty} F_n(z) = \Phi(z) \tag{1}$$

is the following relation for any $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{C_n^2} \sum_{k=1}^{n} \int_{|y-\mu_k| > \varepsilon c_n} (y-\mu_k)^2 dG_k(y) = 0. \tag{2}$$

Instead of formula (2) we will use the following:

$$Z_n \xrightarrow{L} N(0, 1) \tag{3}$$

In sampling scheme with probability proportional to value of characteristic Y with replacement, although general population is finite we can use Lindeberg-Feller theorem. On the grounds of this theorem it can be proved that Hansen-Hurwitz estimator has, see Bracha (1998), with $n \rightarrow \infty$ normal distribution. Let us note that n can be optionally large (sample units can be optionally large and in addition variables Y_i and Y_i' $(i \neq i')$ can be independent (sampling without replacement).

In case when $n < N$ and variates are independent and as a consequence we can not use Lindeberg-Feller theorem to estimate of average value of population.

Difficulties arising from the assumption $n < N$ and variables interdependence as the first tried to solve Madow (1948). He considered, instead given population U, population sequence $\{U_\upsilon\}$, which was generated by multiple reproduction of particular elements from population U, under the assumption that both size of these populations and samples sizes, which are sampled from them tend to infinity, that is by $\upsilon \varkappa \infty$, $N_\upsilon \rightarrow \infty$, $n_\upsilon \rightarrow \infty$ and $\frac{n_\upsilon}{N_\upsilon} \rightarrow q < 1$.

Hajek (1960) reformulated Madow theorem which can be shown as follows:, see also Erdös i Réyi (1959).

Theorem 2 (Lindeberg-Feller-Hajek). There is given a population sequence $\{U_\nu\}_{\nu=1}^{\infty}$, where

$$U_\nu = \{Y_{\nu 1}, \ldots, Y_{\nu N_\nu}\} \tag{4}$$

corresponding sequence $\{Y_\nu\}_{\nu=1}^{\infty}$, where $Y_\nu = (Y_{\nu 1}, \ldots, Y_{\nu N_\nu})^T$, and also data sequence

$$\{d_\nu'\}_{\nu=1}, \quad \text{where} \quad d_\nu' = \{y_{\nu 1}, \ldots, y_{\nu n_\nu}\}$$

and corresponding sequences of general terms

$$\overline{Y}_{\nu} = \frac{1}{N_{\nu}} \sum_{j=1}^{N_{\nu}} Y_{\nu j}, \tag{5}$$

$$S_{\nu}^{2} = \frac{1}{N_{\nu} - 1} (Y_{\nu j} - \overline{Y}_{\nu})^{2}, \tag{6}$$

$$\overline{y}_{\nu} = \frac{1}{n_{\nu}} \sum_{j=1}^{n_{\nu}} y_{\nu i}, \tag{7}$$

Let $U_{\nu\varepsilon}$ be for $\varepsilon > 0$ subset of set U_{ν} and let its elements satisfy a condition

$$|Y_{\nu j} - \overline{Y}_{\nu}| > \varepsilon \sqrt{n_{\nu}\left(1 - \frac{n_{\nu}}{N_{\nu}}\right)} S_{\nu}, \tag{8}$$

when $n_{\nu} \to \infty$, $N\nu \to n_{\nu} \to \infty$ for $\nu = \infty$.
Necessary and sufficient condition to

$$z_{\nu} = \frac{\overline{y}_{\nu} - \overline{Y}_{\nu}}{\sqrt{\left(\frac{1}{n_{\nu}} - \frac{1}{N_{\nu}}\right)} S_{\nu}} \xrightarrow{L} N(0, 1), \tag{9}$$

is relation

$$\lim_{\nu \to \infty} \frac{\sum_{j \in u_{\nu\varepsilon}} (Y_{\nu j} - \overline{Y}_{\nu})^{2}}{\sum_{j \in u_{\nu}} (Y_{\nu j} - \overline{Y}_{\nu})^{2}} = 0 \tag{10}$$

(this relation is called Lindeberga-Hajek condition).

Scott and Wu (1981) proved further features of estimators in case of simple sampling without replacement.

Theorem 3 (Scott-Wu). If the following condition is satisfied

$$\lim_{\nu \to \infty} \left(1 - \frac{n_{\nu}}{N_{\nu}}\right) \frac{S_{\nu}^{2}}{n_{\nu}} = 0, \tag{11}$$

then for $\varepsilon > 0$

$$\lim_{\nu \to \infty} P\{|\overline{y}_{\nu} - \overline{Y}_{\nu}| < \varepsilon\} = 1. \tag{12}$$

Often instead of equality (12) we use the following formula:

$$(\overline{y}_v - \overline{Y}_v) \xrightarrow{P} 0.$$

Theorem 2 shows that average from the sample which was sampled according to simple sample without replacement scheme is compatible with estimator of average $\overline{Y}$.

Hajek (1964) proposed rejective sampling procedure for the scheme of sampling with probabilities proportional to the value characteristic Y without replacement. This scheme proves that for given $p_j = \frac{x_j}{X}$ there are determined by the sizes a_j being function p_j and fulfilling a condition

$$\sum_{j=1}^{N} a_j = 1$$

Next, units are sampled with replacement and choice probabilities in each drawing proportional to a_j. If sample contains n various units, then the sample is accepted. But if some units repeat then the whole sample is rejected and the new sample is sampled.

Rosén (1972) proved central limit theorem for Horvitz-Thompson estimator based on random sequence sample. These analyses are performed by many researchers by means of both analytic and simulated methods, see for example Bracha (1990; 1998). On the grounds of results of these research authors suggest high caution in drawing conclusions about distribution compatibility of considered estimators with normal distribution.

III. χ^2 TEST OF GOODNESS OF FIT FOR COMPLEX SAMPLES

Let variate X take values belonging to $k(k \geqslant 2)$ separable intervals. Let us denote by p_i probability that variable **X** takes values from i-value interval and at the same time $p_i > 0$ for $i = 1, ..., k$ and $\sum_{i=1}^{n} p_i = 1$. On the grounds of simple sample the hypothesis must be verified:

$$H_0 : \mathbf{p} = \mathbf{p}_0$$

towards to alternative hypothesis:

$$H_0 : \mathbf{p} \neq \mathbf{p}_0,$$

where: $\mathbf{p} = [p_i]_{i=1,...,k-1}$, $\mathbf{p}_0$ is $(k-1)$ dimensional vector of hypothetical probabilities connected with $\mathbf{p}(\mathbf{p}_0 = [p_{0i}]_{i=1,...,k-1})$.

To verify hypothesis H_0 it is proposed to use matrix statistics, see for example Rao (1982).

$$\chi^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)^T \mathbf{p}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0), \tag{13}$$

where:

$$P_0 = diag(p_0) - \mathbf{p}_0\mathbf{p}_0^T, \quad \hat{p} = [\hat{p}_i]_{i=1,\dots,k-1} \tag{14}$$

and at the same time $\hat{p}_i$ is unbiased estimator p_i.

Under the assumption that veracity of hypothesis H_0 statistics given by formula (13) has asymptotic distribution χ^2 of $k-1$ degrees of freedom.

For complex samples Holt and others (1980) showed modifications of χ^2 goodness of fit statistics which has the following form:

$$\chi_*^2 = \frac{\chi^2}{\hat{\lambda}} \tag{15}$$

where:

$$\hat{\lambda} = \frac{n}{k-1} \sum_{i=1}^{k} \frac{\hat{\mathrm{D}}^2(p_i)}{p_{io}} \tag{16}$$

and at the same time $\hat{\mathrm{D}}(p_i)$ denote variance estimators of investigated characteristic which are suitable for particular sampling scheme. Taking into account a variance of hypothesis H_0 statistics (15) has χ^2 distribution of $(k-1)$ degrees of freedom. We reject hypothesis H_0 on the significance level α, if inequality $\chi_*^2 \geqslant \chi_\alpha^2$ proceeds.

In case when $k = 2$, we verify hypothesis $H_0: p = p_0$ against alternative hypothesis $H_1: p \neq p_0$ by means of statistics, see for example Bracha (1998).

$$\chi_*^2 = \frac{(\hat{p} - p_0)^2}{\hat{\mathrm{D}}^2(p)} \tag{17}$$

where $\hat{p}$ is estimator p.

Statistics (17) by the veracity of hypothesis H_0 has for big values n distribution close to distribution χ^2 of one goodness of fit.

We made a few experiments using Monte Carlo method for complex samples investigating sizes of χ^2 test and its modification χ_*^2. In the first experiment we were comparing sizes of investigated tests for complex samples (non-returnable sampling) in finite population of normal distribution with demanded parameters for $N = 1000$, 2000, 10 000. On the ground of sampled samples we were verifying simple hypothesis H_0, that sample comes from population of normal distribution by means of classic test χ^2 and modified χ^2_* taking into consideration sampling scheme effect. The investigation was made for dozen or so variants of classes division of sample results for example number of classes. $N = 1000$ $k = 4, 5, 6, 8, 10, 12, 15, 20$ adequately to size of sample which fulfils conditions of convergence statistics χ^2 towards distribution χ^2, see Domański (1990). The investigation was made for $q = 10\ 000$ repetitions.

In Table 1 we illustrated sizes of considered tests for three significance levels $\alpha = 0.10$; 0.05; 0.01 and number of degrees of freedom ($lss = 7$) for $N = 1000$, 2000 and $lss = 14$ for $N = 10\ 000$. On the contrary in Table 2 for $N = 1000$ we presented considered tests sizes for ($lss = 2, 4, 6$) depending on number of degrees of freedom.

IV. CONCLUSIONS

1. The size of test χ^2 for $N = 1000$ in all cases exceeds assumed significance levels and on the contrary modified test χ^2_* does not exceed assumpted significance levels $\alpha = 0.10$ and $\alpha = 0.05$, and also generally for $\alpha = 0.01$. We obtained similar results for $N = 10\ 000$ (see Table 1).

Table 1. Comparison of size of χ^2 goodness of fit with modified test χ^2 for complex samples sampled from finite populations of normal distribution for $N = 1000$, 2000, 10 000 $lss = (k-1)$ degrees of freedom

n sample size	Significance level					
	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	χ^2	χ^2_*	χ^2	χ^2_*	χ^2	χ^2_*
$N = 1000$ ($lss = 7$)						
40	0.136	0.091	0.071	0.046	0.020	0.010
50	0.128	0.086	0.071	0.046	0.023	0.012
60	0.123	0.085	0.068	0.047	0.018	0.008
70	0.111	0.078	0.067	0.040	0.020	0.013
80	0.105	0.081	0.064	0.042	0.017	0.013
90	0.116	0.090	0.057	0.041	0.014	0.009
100	0.106	0.092	0.062	0.053	0.014	0.010
120	0.111	0.090	0.059	0.048	0.016	0.008

Table 1. (contd.)

n sample size	Significance level					
	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	χ^2	χ^2_*	χ^2	χ^2_*	χ^2	χ^2_*
$N = 2000$ ($lss = 7$)						
50	0.128	0.089	0.064	0.042	0.021	0.012
100	0.102	0.071	0.057	0.034	0.015	0.009
150	0.097	0.082	0.054	0.043	0.010	0.007
200	0.076	0.057	0.033	0.025	0.005	0.005
300	0.083	0.087	0.051	0.052	0.009	0.010
$N = 10\ 000$ ($lss = 14$)						
200	0.134	0.104	0.081	0.062	0.024	0.012
300	0.116	0.088	0.068	0.053	0.021	0.014
400	0.125	0.106	0.075	0.063	0.010	0.007
500	0.103	0.097	0.049	0.046	0.014	0.009

Source: Own calculations.

2. With the increase of number of degrees of freedom in general size of classic test χ^2 more and more stands off obtained significance level and, on the contrary, with the increase of number of freedom, size of modified tests χ^2_* more and more approach the assumed significance level (see Table 2).

Table 2. Comparison of size of χ^2 goodness of fit with modified test χ^2_* for complex samples sampled from finite populations of normal distribution for $N = 1000$ depending on number of degrees of freedom $lss = 2, 4, 6$

n sample size	Significance level					
	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	χ^2	χ^2_*	χ^2	χ^2_*	χ^2	χ^2_*
$lss = 2$						
10	0.1145	0.0593	0.0670	0.0318	0.0192	0.0093
15	0.1095	0.0490	0.0612	0.0259	0.0160	0.0070
20	0.1026	0.0497	0.0540	0.0221	0.0147	0.0064
30	0.0979	0.0427	0.0502	0.0202	0.0119	0.0046
40	0.0871	0.0375	0.0441	0.0188	0.0104	0.0040
50	0.0857	0.0379	0.0425	0.0178	0.0084	0.0036
100	0.0722	0.0367	0.0319	0.0163	0.0058	0.0027
$lss = 4$						
15	0.1398	0.0831	0.0813	0.0452	0.0263	0.0128
20	0.1272	0.0756	0.0737	0.0406	0.0224	0.0105

Table 2. (contd.)

n sample size	Significance level					
	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	χ^2	χ^2_*	χ^2	χ^2_*	χ^2	χ^2_*
$lss = 4$						
30	0.1204	0.0730	0.0699	0.0368	0.0215	0.0096
40	0.1237	0.0768	0.0682	0.0384	0.0208	0.0101
50	0.1163	0.0715	0.0651	0.0408	0.0187	0.0105
100	0.1056	0.0727	0.0533	0.0367	0.0139	0.0082
$lss = 6$						
20	0.1516	0.0982	0.0906	0.0533	0.0331	0.0167
30	0.1405	0.0930	0.0851	0.0506	0.0285	0.0143
40	0.1327	0.0883	0.0779	0.0492	0.0249	0.0138
50	0.1213	0.0857	0.0727	0.0463	0.0222	0.0119
100	0.1161	0.0934	0.0625	0.0492	0.0160	0.0122

Source: Own calculations.

Summing up, it has to be emphasised that on this stage for complex sample (dependent sampling) classic χ^2 test of goodness of fit in general gives in assumed cases insatiable indications in relation to hypothesis verification. Most often in sampling without replacement real error of the first type considerably exceeds obtained significance level α.

From the experience gathered so far it follows that assumed test should be investigated for simple and complex samples. Therefore, some postulates of many authors who refer to rules of applying χ^2 should be verified.

REFERENCES

Bracha Cz. (1990), *Wybrane problemy wnioskowania statystycznego na podstawie prób nieprostych*, ZBSE GUS, PAN, Warszawa.

Bracha Cz. (1998), *Metoda reprezentacyjna w badaniach opinii publicznej i marketingu*, EFEICT, Warszawa.

Domański Cz. (1990), *Testy statystyczne*, PWE, Warszawa.

Erdös P., Réyi A. (1959), On the central limit theorem for samples from a finite population, *Publications of the Mathematics Institute of Hungarian Academy of Science*, **4**, 49–57.

Fisz M. (1967), *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN, Warszawa.

Hajek J. (1960), Limiting distribution in sample random sampling from a finite populations, *Publications of the Matematics Institute of the Hungarian Academy of Science*, **5**, 361–374.

Hajek J (1964), Asymptotic theory of rejective sampling with varying probabilities from a finite population, *Annals of Mathematical Statistics*, 1491–1523.

Holt D., Scott A.J., Evings P.D. (1980), Enings chi-squared test with survey, *Journal of the American Statistical Association*, Ser. A, **143**, 303–320.

Madow W.G. (1948), On the limiting distributions of estimates based on sample from finite populations, *Annals of Mathematical Statistics*, **19**, 535–545.

Rao C.K. (1982), *Modele liniowe statystyki matematycznej*, PWN, Warszawa.

Rosén B. (1972), Asymptotic theory for successive sampling with varying probabilities without replacement: I and II, *Annals of Mathematical Statistics*, **43**, 373–397 and 748–776.

Scott A. J., Wu C. F. (1981), On the asymptotic distributions of ratio and regression estimator, *Journal of American Statistical Association*, **76**, 98–102.

Czesław Domański

UWAGI O WNIOSKOWANIU STATYSTYCZNYM DLA PRÓB NIEPROSTYCH

Streszczenie

Klasyczna teoria wnioskowania statystycznego dostarcza nam metod estymacji nieznanych parametrów rozkładu, szacowanie postaci funkcji określającej ten rozkład oraz weryfikację hipotez na podstawie prób prostych, tzn. takich, w których obserwacje są niezależne i mają ten sam rozkład prawdopodobieństwa. Na ogół jednak ze względu na koszty i efektywność badań posługujemy się próbami nieprostymi lub złożonymi (*complex samples*). Wyniki obserwacji w tych próbach są realizacjami stochastycznie zależnych zmiennych losowych o różnych rozkładach. W badaniach reprezentacyjnych wyróżniamy między innymi następujące schematy: losowanie zależne (bez zwracania), losowanie z różnymi prawdopodobieństwami wyboru, warstwowe, zespołowe i wielostopniowe. Przykładowo, losowanie bez zwracania eliminuje stochastyczną niezależność obserwacji, proces warstwowania zróżnicowanie prawdopodobieństw wyboru elementów próby, natomiast losowanie wielostopniowe wpływa na różnorodność rozkładów.

Przedmiotem tej pracy są problemy związane z estymacją (metody adaptacji centralnego twierdzenia granicznego dla prób nieprostych) oraz weryfikacja hipotez o zgodności rozkładów dla prób nieprostych.