Analysis of samples from finite population

Jurgita TURKUVIENĖ, Algimantas BIKELIS (VU)
e-mail: jurgutet23@yahoo.com, marius@post.omnitel.net

Abstract. This paper deals with asymptotic properties of probability distributions of sample statistics when samples are selected from finite populations. These properties also were analysed by P. Erdős, A. Rényi [3] and J. Hájek [4]. Relationships between probability distributions of sample sums were investigated in [6] article.

Keywords: finite population, Appel polynomial, Bernoulli sampling.

1. Introduction

Assume that probability sample is selected from finite population

$$\mathcal{O} = \{O_1, O_2, \dots, O_N\}.$$

Sample statistics are defined in finite probability space $\{\Omega, \mathcal{A}, P\}$ therefore their distributions are quasi-lattice when elements of population $\mathcal{O}$ are characterized by sequence of real numbers

$$a_{1\nu}, a_{2\nu}, \dots, a_{N_\nu \nu}, \quad \nu = 1, 2, \dots. \tag{1}$$

J. Hájek analyzed simple random sampling and Bernoulli sampling from this sequence. Our study covers not only these two samplings, but also simple random sampling with replacement.

It is known [2] that integer base $\vec{\beta}_\nu = (\beta_{1\nu}, \beta_{2\nu}, \dots, \beta_{k_\nu \nu})$ exists in finite sequence of real numbers (1). This base is such that $\vec{\beta}_{jl} > 0$ and

$$a_{j\nu} = (\vec{b}_\nu, \vec{E}) + (\vec{\beta}_\nu, \vec{m}_j), \tag{2}$$

where $\vec{b}_\nu \in R^{k_\nu}$, $\vec{E} = (1, 1, \dots, 1) \in R^{k_\nu}$, $(\vec{b}_\nu, \vec{E})$ is scalar product, $\vec{m}_j = (m_{1j}, m_{2j}, \dots, m_{k_\nu j})$, $m_{ij} = 0, \pm 1, \pm 2, \dots$.

It should be mentioned that dimension k_ν of space R^{k_ν} depends on series number ν and can increase when $\nu \to \infty$. When base is selected, representation (2) is univalent. So sequence (1) can be replaced with sequence

$$(\vec{b}_\nu, \vec{E}) + (\vec{\beta}_\nu, \vec{m}_1), \dots, (\vec{b}_\nu, \vec{E}) + (\vec{\beta}_\nu, \vec{m}_{N_\nu}).$$

We will omit number ν of series, i.e., we will analyze $a_j = (\vec{b}, \vec{E}) + (\vec{\beta}, \vec{m}_j)$, $j = 1, 2, \dots, N$.

When simple random sample $(\vec{b},\vec{E})+(\vec{\beta},\ \vec{m}_{i_1})$, $(\vec{b},\vec{E})+(\vec{\beta}\vec{m}_{i_2}),\ldots,$ $(\vec{b},\vec{E})+(\vec{\beta},\vec{m}_{i_n})$, $1\leqslant i_1\leqslant\ldots\leqslant i_n\leqslant N$, is selected from finite population, sample sum $S_{nN}=\sum_{j=1}^{n}((\vec{b},\vec{E})+(\vec{\beta},\vec{m}_{i_j}))$ obtains values $n(\vec{b},\vec{E})+(\vec{\beta},\vec{m}_{i_j})$, where $\vec{\nu}_{i_j}=(\nu_{1i_j},\ldots,\nu_{n_k i_j})$, $\nu_{lr}=0,\pm1,\pm2,\ldots$.

P. Erdős and A. Rényi [3] proved that characteristic function of S_{nN} is

$$\mathrm{M}\mathrm{e}^{itS_{nN}}=\frac{1}{C_N^n}\sum_{1\leqslant i_1\leqslant\ldots\leqslant i_n\leqslant N}\mathrm{e}^{it(a_{i_1}+\ldots+a_{i_n})}$$
$$=\frac{1}{2\pi P_N(n)}\int_{-\pi}^{pi}\mathrm{e}^{-i\theta n}\prod_{j=1}^{N}\left(q+p\mathrm{e}^{i\theta+ita_j}\right)\mathrm{d}\theta,$$

where $P_N(n)=C_N^n p^n q^{N-n}$, $p=\frac{n}{N}$.

J. Hájek [4] analyzed sum $S_{\mu N}=\sum_{j=1}^{\mu}a_{i_j}$ of Bernoulli sample $a_{i_1},a_{i_2},\ldots,a_{i_\mu}$, $i_1\neq\ldots\neq i_\mu$, where distribution of random number μ is Binomial distribution $B(N,p)$.

He showed that $\mathrm{M}\mathrm{e}^{itS_{\mu N}}=\prod_{j=1}^{N}(q+p\mathrm{e}^{ita_j})$. So

$$\mathrm{M}\mathrm{e}^{itS_{\mu N}}=\prod_{j=1}^{N}\left(q+p\mathrm{e}^{it((\vec{b},\vec{E})+(\vec{\beta},\vec{m}_j))}\right).$$

2. Coefficient of correlation

Denote $\vec{\eta}_1,\vec{\eta}_2,\ldots,\vec{\eta}_N$ – k-dimensional independent random vectors which obtain values $\vec{0}$ and $\vec{b}+\vec{\beta}\vec{m}_j$, $j=1,2,\ldots,N$ with probabilities q and p, i.e., $P\{\vec{\eta}=\vec{0}\}=q$ and $P\{\vec{\eta}=\vec{b}+\vec{\beta}\vec{m}_j\}=p$, where $\vec{\beta}\vec{m}_j=(\beta_1 m_{1j},\beta_2 m_{2j},\ldots,\beta_k m_{kj})$. Let $\vec{Z}_{nN}=\sum_{j=1}^{N}\vec{\eta}_j=(Z_1,Z_2,\ldots,Z_k)$. We are interested in correlation matrix of this vector when

$$\mathrm{M}\mathrm{e}^{i(\vec{t},\vec{Z}_{nl})}=\prod_{j=1}^{N}\left(q+p\mathrm{e}^{i(\vec{t},\vec{b})+i(\vec{t},\vec{\beta}\vec{m}_j)}\right).$$

We obtain $\vec{Z}_{nN}=\sum_{j=1}^{N}\vec{\eta}_j=(Z_1,Z_2,\ldots,Z_k)$, $\mathrm{D}Z_l=pq(b_l+\beta_l m_l)^2$ and

$$\mathrm{M}\big[(Z_l-\mathrm{M}Z_l)(Z_r-\mathrm{M}Z_r)\big]=pq\sum_{j=1}^{N}(b_l+\beta_l m_{lj})(b_r+\beta_r m_{rj}).$$

Denote coefficient of correlation of random variables Z_l and Z_r by ρ_{lr}. Then we have

$$\rho_{rl}=\frac{\sum_{j=1}^{N}(b_l+\beta_l m_{lj})(b_r+\beta_r m_{rj})}{\left(\sum_{j=1}^{N}(b_l+\beta_l m_{lj})^2\sum_{i=1}^{N}(b_r+\beta_r m_{ri})^2\right)^{\frac{1}{2}}}.$$

It is known from [1], that Cauchy inequality

$$\Big(\sum_{j=1}^{N} x_j y_j\Big)^2 \leqslant \Big(\sum_{j=1}^{N} x_j^2\Big)\Big(\sum_{i=1}^{N} y_j^2\Big)$$

becomes equality if and only if when there exist numbers λ and μ such that $\lambda x_j + \mu y_j = 0$ for all $j = 1, 2, \ldots, N$ and λ and μ are not equal to 0 at the same time.

LEMMA 1. *Coefficient of correlation ρ_{lr} is equal to ± 1, when there exist λ and μ not equal to zero at the same time such that*

$$\lambda b_l + \mu b_r + \lambda \beta_l m_{lj} + \mu \beta_r m_{rj} = 0.$$

When $b_l = b_r = 0$, then necessary and sufficient condition for $\rho_{rl} = \pm 1$ is $\lambda m_{lj} + \mu m_{rj} = 0$, $j = 1, 2, \ldots, N$.

This property is very important when probabilities

$$P\big\{S_{\mu n} = l(\vec{b}, \vec{E}) + (\vec{\beta}, \vec{\nu})\big\} = P\big\{\vec{Z}_{nN} = l\vec{b} + \vec{\beta}\vec{\nu}\big\}$$
$$= \frac{\beta_1, \beta_2, \ldots, \beta_k}{(2\pi)^k} \int_{-\frac{\pi}{\beta_1}}^{\frac{\pi}{\beta_1}} \cdots \int_{-\frac{\pi}{\beta_k}}^{\frac{\pi}{\beta_k}} \mathrm{e}^{-i(\vec{t}, l\vec{b}) - i(\vec{t}, \vec{\beta}\vec{\nu})} \mathrm{M}\mathrm{e}^{i(\vec{t}, \vec{Z}_{nN})} \, \mathrm{d}\vec{t}.$$

are analyzed.

Notice that random vector $\vec{Z}_{nN} = \sum_{j=1}^{N} \vec{\eta}_j$ is a sum of independent differently distributed random vectors $\vec{\eta}_1, \vec{\eta}_2, \ldots, \vec{\eta}_N$. It is possible to analyze sums of identically distributed random vectors instead of sums of differently distributed random vectors. Two ways of such analysis are known for us. The first way is to use characteristic functions of accompanying distributions. Second way is explained more detailed below.

3. Formal expansion

Denote k-dimensional characteristic functions by $f_1, f_2, \ldots, f_N$ and their sum $g = \frac{1}{N} \sum_{j=1}^{N} f_j$. Let $T = \{\vec{t}\colon\ f_j(\vec{t}) \neq 0,\ j = 1, 2, \ldots, N \text{ and } \vec{t} \in R^k\}$.

Then, if $\vec{t} \in T$ and $|\frac{f_j - g}{g}| \leqslant C < 1$ we have

$$\prod_{j=1}^{N} f_j = g^N \exp\bigg\{\sum_{j=1}^{N} \frac{f_j - g}{g} - \frac{1}{2}\sum_{j=1}^{N}\Big(\frac{f_j - g}{g}\Big)^2 + \sum_{j=1}^{N}\sum_{m=3}^{\infty} \frac{(-1)^{m-1}}{m}\Big(\frac{f_j - g}{g}\Big)^m\bigg\}.$$

Here for all $\vec{t} \in R^k$ $\sum_{j=1}^{N} \frac{f_j - g}{g} = 0$. We get

$$\prod_{j=1}^{N} f_j = g^N \exp\bigg\{-\frac{1}{2}\frac{1}{N}\sum_{j=1}^{N}\Big(\sqrt{N}\frac{f_j - g}{g}\Big)^2\bigg\}\bigg[1 + \sum_{j=1}^{\infty} \beta_j \Big(\frac{1}{\sqrt{N}}\Big)^j\bigg],$$

where

$$\beta_j = \sum_{\nu_1+2\nu_2+\cdots+j\nu_j=j} \frac{\alpha_1^{\nu_1}\ldots\alpha_j^{\nu_j}}{\nu_1!\ldots\nu_j!},$$

$$\alpha_l = \frac{(-1)^{l+1}}{l+2}\frac{1}{N}\sum_{j=1}^{N}\Big(\sqrt{N}\frac{f_j-g}{g}\Big)^{l+2}.$$

From these equations it is seen that powers of $\frac{f_j-g}{g}$ are used for expansion, e.g.,

$$\Big(\sqrt{N}\frac{f_j-g}{g}\Big)^m = \bigg(\frac{1}{N}\sum_{i=1}^{N}\Big(\sqrt{N}\frac{f_j-f_i}{g}\Big)\bigg)^m$$

and

$$\frac{1}{N}\sum_{j=1}^{N}\Big(\sqrt{N}\frac{f_j-g}{g}\Big)^m = \frac{1}{N}\sum_{j=1}^{N}\bigg(\frac{1}{N}\sum_{i=1}^{N}\Big(\sqrt{N}\frac{f_j-f_i}{g}\Big)\bigg)^m.$$

Further characteristic function h^N of known distribution is used to approximate

$$g^N = \bigg(\frac{1}{N}\sum_{j=1}^{N} f_j\bigg)^N.$$

When $\sqrt{N}|\frac{g-h}{g}| < 1$ we obtain

$$g^N = h^N \exp\bigg\{N\frac{g-h}{h} - \frac{1}{2}\Big(\sqrt{N}\frac{g-h}{h}\Big)^2\bigg\}\bigg[1+\sum_{j=1}^{\infty}\Big(\frac{1}{\sqrt{N}}\Big)^j A_{j2}\Big(\sqrt{N}\frac{g-h}{h}\Big)\bigg].$$

Here $A_{j2}\big(\sqrt{N}\frac{g-h}{h}\big)$ is generalized Appel polynomial. It is known [5], that

$$A_{j2}(y) = (-1)^{j+1}y^{j+2}\sum_{k=0}^{j-1}(-1)^{3k}g_{jk}^{(2)}y^{2k},$$

$$g_{jk}^{(2)} = \sum_{\substack{\nu_1+2\nu_2+\cdots+j\nu_j=j\\ \nu_1+\nu_2+\cdots\nu_j=k}} \frac{1}{\nu_1!\nu_2!\ldots\nu_j!}\Big(\frac{1}{3}\Big)^{\nu_1}\cdots\Big(\frac{1}{j+2}\Big)^{\nu_j}.$$

Now

$$\prod_{j=1}^{N} f_j = h^N \exp\bigg\{N\frac{g-h}{h} - \frac{1}{2}\Big(\sqrt{N}\frac{g-h}{h}\Big)^2 - \frac{1}{2}\frac{1}{N}\sum_{j=1}^{N}\Big(\sqrt{N}\frac{f_j-g}{g}\Big)^2\bigg\}$$

$$\times\left[1+\sum_{j=1}^{\infty}\Bigl(\frac{1}{\sqrt{N}}\Bigr)^{j}\beta_j][1+\sum_{j=1}^{\infty}\Bigl(\frac{1}{\sqrt{N}}\Bigr)^{j}A_{j2}\Bigl(\sqrt{N}\frac{g-h}{h}\Bigr)\right].$$

This formal expansion is valid, because values of functions $\frac{f_j-g}{g}$ and and $\frac{g-h}{h}$ are close to 0 in the neighborhood of point $\vec{t}=\vec{0}\in R^k$. Conditions $\sqrt{N}|\frac{f_j-g}{g}|<1$ and $\sqrt{N}|\frac{g-h}{h}|<1$ are sufficient for convergence of series. *Remark.* All sums of samples from finite populations are sums of differently distributed random numbers or variables. Terms of sums can be dependent or not.

We artificially used factor $\sqrt{N}$, i.e., we took $\sqrt{N}\frac{f_j-g}{g}$ and $\sqrt{N}\frac{g-h}{h}$ because characteristic function h from infinitely divisible distributions subset $\mathcal{L}$ will be used in our further studies. When samples are selected from finite populations, terms of sample sums are random numbers or vectors having finite moments of all orders. It is important that matrix of the second order moments has to to be non-singular. If conditions of lemma are fulfilled then matrix is definitely non-singular.

References

1. E.F. Bechenbach, R. Bellman, *Inequalities*, Springer Verlag, Berlin (1962).
2. A. Bikelis, Asymptotic expansions for distribution of statistics, in: *Proceedings of the XXXVI Conference of Lithuanian Math. Soc.* (1996), pp. 5–28.
3. P. Erdös, A. Rényi, On a central limit theorem for samples from a finite population, *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **4**, 49–61 (1959).
4. J. Hájek, Limiting distributions in simple random sampling for a finite population, *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **5**, 361–374 (1960).
5. V.M. Kalinin, *Special functions and limit properties of probability distributions*, *Notes from Scientific Seminars, LOMI*, **13**, 5–137 (1968) (in Russian).
6. J. Turkuvienė, A. Bikelis, On the distribution of sample mean in finite populations, *Lith. Math. J.*, **48**(1), 100–122 (2008).

REZIUMĖ

J. Turkuvienė, A. Bikelis. Imčių ir baigtinių visumų analizė

Straipsnyje nagrinėjamos asimptotinės imčių iš baigtinių visumų tikimybinių skirstinių savybės.

Raktiniai žodžiai: baigtinė populiacija, Apelio polinomai, Bernulio imtis.