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

\[ \mathcal{O} = \{ O_1, O_2, \ldots, O_N \}. \]

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

\[ a_{1v}, a_{2v}, \ldots, a_{Nv}, \quad v = 1, 2, \ldots \] (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 \( \vec{\beta}_v = (\beta_{1v}, \beta_{2v}, \ldots, \beta_{kv}) \) exists in finite sequence of real numbers (1). This base is such that \( \vec{\beta}_{jl} > 0 \) and

\[ a_{jv} = (\vec{b}_v, \vec{E}) + (\vec{\beta}_v, \vec{m}_j) \] (2)

where \( \vec{b}_v \in \mathbb{R}^k, \vec{E} = (1, 1, \ldots, 1) \in \mathbb{R}^k, (\vec{b}_v, \vec{E}) \) is scalar product, \( \vec{m}_j = (m_{1j}, m_{2j}, \ldots, m_{kj}), m_{ij} = 0, \pm 1, \pm 2, \ldots \)

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

\[ (\vec{b}_v, \vec{E}) + (\vec{\beta}_v, \vec{m}_1), \ldots, (\vec{b}_v, \vec{E}) + (\vec{\beta}_v, \vec{m}_{Nv}). \]

We will omit number \( v \) of series, i.e., we will analyze \( a_j = (\vec{b}, \vec{E}) + (\vec{\beta}, \vec{m}_j), j = 1, 2, \ldots, N. \)
When simple random sample \((\vec{b}, \vec{E}) + (\vec{\beta}, \vec{m}_i), \, 1 \leq i_1 \leq \ldots \leq i_n \leq N\), is selected from finite population, sample sum \(S_{nN} = \sum_{j=1}^n ((\vec{b}, \vec{E}) + (\vec{\beta}, \vec{m}_i))\) obtains values \(n(\vec{b}, \vec{E}) + (\vec{\beta}, \vec{m}_i)\), where \(\vec{v}_{ij} = (v_{1ij}, \ldots, v_{nij})\), \(v_{fr} = 0, \pm 1, \pm 2, \ldots\).

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

\[
\Me^{itS_{nN}} = \frac{1}{C_N^n} \sum_{1 \leq i_1 \leq \ldots \leq i_n \leq N} e^{it(a_{i_1} + \ldots + a_{i_n})}
\]

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

J. Hájek [4] analyzed sum \(S_{\mu N} = \sum_{j=1}^\mu a_{ij}\) 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 \(\mu\) is Binomial distribution \(B(N, p)\).

He showed that \(\Me^{itS_{\mu N}} = \prod_{j=1}^N (q + p e^{it(\vec{b}, \vec{E}) + (\vec{\beta}, \vec{m}_j)})\). So

\[
\Me^{itS_{\mu N}} = \prod_{j=1}^N (q + p e^{it((\vec{b}, \vec{E}) + (\vec{\beta}, \vec{m}_j)))}.
\]

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., \(P(\vec{\eta} = 0) = q\) 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

\[
\Me^{i(\vec{\eta}, \vec{Z}_n)} = \prod_{j=1}^N (q + pe^{i((\vec{b}, \vec{E}) + (\vec{\beta}, \vec{m}_j)))}.
\]

We obtain \(\vec{Z}_nN = \sum_{j=1}^N \vec{\eta}_j = (Z_1, Z_2, \ldots, Z_k), \, DZ_l = pq(b_l + \beta_m)^2\) and

\[
M[(Z_l - MZ_l)(Z_r - MZ_r)] = pq \sum_{j=1}^N (b_l + \beta m_{lj})(b_r + \beta m_{rj}).
\]

Denote coefficient of correlation of random variables \(Z_l\) and \(Z_r\) by \(\rho_{lr}\). Then we have

\[
\rho_{lr} = \frac{\sum_{j=1}^N (b_l + \beta m_{lj})(b_r + \beta m_{rj})}{(\sum_{j=1}^N (b_l + \beta m_{lj})^2 \sum_{j=1}^N (b_r + \beta m_{rj})^2)^{\frac{1}{2}}}.
\]
It is known from [1], that Cauchy inequality
\[
\left( \sum_{j=1}^{N} x_j y_j \right)^2 \leq \left( \sum_{j=1}^{N} x_j^2 \right) \left( \sum_{i=1}^{N} y_i^2 \right)
\]
becomes equality if and only if when there exist numbers \( \lambda \) and \( \mu \) such that \( \lambda x_j + \mu y_j = 0 \) for all \( j = 1, 2, \ldots, N \) and \( \lambda \) and \( \mu \) are not equal to 0 at the same time.

**Lemma 1.** Coefficient of correlation \( \rho_{ij} \) is equal to \( \pm 1 \), when there exist \( \lambda \) and \( \mu \) not equal to zero at the same time such that
\[
\lambda b_i + \mu b_r + \lambda \beta_i m_{ij} + \mu \beta_r m_{rj} = 0.
\]
When \( b_i = b_r = 0 \), then necessary and sufficient condition for \( \rho_{ij} = \pm 1 \) is \( \lambda m_{ij} + \mu m_{rj} = 0, j = 1, 2, \ldots, N \).

This property is very important when probabilities
\[
P\left\{ \sum_{j=1}^{N} \eta_j = l(\vec{b}, \vec{E}) + (\vec{\beta}, \vec{\nu}) \right\} = P\left\{ \vec{Z}_{\eta N} = \vec{l} \right\}
\]
are analyzed.

Notice that random vector \( \vec{Z}_{\eta N} = \sum_{j=1}^{N} \vec{\eta}_j \) 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 \( T = \{ \vec{r} : f_j(\vec{r}) \neq 0, j = 1, 2, \ldots, N \} \) and \( \vec{r} \in R^k \).

Then, if \( \vec{r} \in T \) and \( |\vec{r}| \leq C < 1 \) we have
\[
\prod_{j=1}^{N} f_j = g^N \exp \left\{ \frac{N}{2} \sum_{j=1}^{N} \left( \frac{f_j - g}{g} \right)^2 + \sum_{j=1}^{N} \sum_{m=3}^{\infty} \frac{(-1)^{m-1}}{m} \left( \frac{f_j - g}{g} \right)^m \right\}.
\]

Here for all \( \vec{r} \in R^k \) \( \sum_{j=1}^{N} \frac{f_j - g}{g} = 0 \). We get
\[
\prod_{j=1}^{N} f_j = g^N \exp \left\{ \left[ 1 + \sum_{j=1}^{\infty} \beta_j \left( \frac{1}{g} \right)^j \right] \right\}.
\]
where

\[ \beta_j = \sum_{v_1 + 2v_2 + \cdots + jv_j = j} \frac{a_1^{v_1} \cdots a_j^{v_j}}{v_1! \cdots v_j!} \]

\[ \alpha_l = (-1)^{l+1} \frac{1}{l+2} \frac{1}{N} \sum_{j=1}^{N} \left( \sqrt{N} f_j - g \right)^{l+2} \]

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

\[ \left( \frac{\sqrt{N} f_j - g}{g} \right)^m = \left( \frac{1}{N} \sum_{i=1}^{N} \left( \sqrt{N} f_j - f_i \right) \right)^m \]

and

\[ \frac{1}{N} \sum_{j=1}^{N} \left( \sqrt{N} f_j - g \right)^m = \frac{1}{N} \sum_{j=1}^{N} \left( \frac{1}{N} \sum_{i=1}^{N} \left( \sqrt{N} f_j - f_i \right) \right)^m \]

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

\[ g^N = \left( \frac{1}{N} \sum_{j=1}^{N} f_j \right)^N \]

When \( \sqrt{N} \frac{g-h}{h} < 1 \) we obtain

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

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

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

\[ g_{j2}(2) = \sum_{v_1 + 2v_2 + \cdots + jv_j = j} \frac{1}{v_1! \cdots v_j!} \left( \frac{1}{3} \right)^{v_1} \cdots \left( \frac{1}{j+2} \right)^{v_j} \]

Now

\[ \prod_{j=1}^{N} f_j = h^N \exp \left\{ N \frac{g-h}{h} - \frac{1}{2} \left( \sqrt{N} \frac{g-h}{h} \right)^2 - \frac{1}{2} \frac{1}{N} \sum_{j=1}^{N} \left( \sqrt{N} \frac{f_j - g}{g} \right)^2 \right\} \]
This formal expansion is valid, because values of functions \( \frac{f_j - g}{g} \) and \( \frac{g - h}{h} \) are close to 0 in the neighborhood of point \( \vec{t} = \vec{0} \in \mathbb{R}^k \). 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 \( L \) 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