A GOODNESS-OF-FIT TEST FOR UNIFORM DISTRIBUTION ON A FINITE GROUP

V.I. Kruglov
Steklov Mathematical Institute, Russian Academy of Sciences
Moscow, RUSSIA
e-mail: kruglov@mi.ras.ru

Abstract

Equiprobable samples with replacements from finite Abelian groups are considered. Limit theorems are proved describing convergence of the distribution of the number of ordered subsamples meeting specified linear relations to Poisson distributions. Basing on these theorems we construct a goodness-of-fit test which checks the hypothesis on the uniform distribution of sample elements.

1 Linear relations for elements of Abelian groups

Let k be a fixed natural number and $a_1, \ldots, a_k$ be fixed non-zero integers. Denote by $w(a_1, \ldots, a_k)$ the number of permutations $\sigma \in S_k$ such that the sets of integer-valued solutions of the equations

$$a_1 y_1 + a_2 y_2 + \ldots + a_k y_k = 0 \qquad (*)$$

and

$$a_{\sigma(1)} y_1 + a_{\sigma(2)} y_2 + \ldots + a_{\sigma(k)} y_k = 0$$

coincide. It is easy to see that equation $(*)$ can be rewritten as follows:

$$a'_1 y_1 + \ldots + a'_1 y_{k_1} + a'_2 y_{k_1+1} + \ldots + a'_2 y_{k_1+k_2} + \ldots + a'_l y_{k_1+\ldots+k_l} = 0,$$

where $k_1 + \ldots + k_l = k$ and $a'_u \neq a'_v$ for $u \neq v$.

Lemma 1. *If there exists a permutation $\sigma \in S_k$ such that*

$$(-a_1, \ldots, -a_k) = \left(a_{\sigma(1)}, \ldots, a_{\sigma(k)}\right),$$

then $w(a_1, \ldots, a_k) = 2 \cdot k_1! \cdot k_2! \cdot \ldots \cdot k_l!$, otherwise $w(a_1, \ldots, a_k) = k_1! \cdot k_2! \cdot \ldots \cdot k_l!$.

Let G be a finite additive Abelian group and let $x_1, \ldots, x_T$ be independent random variables uniformly distributed on G. Denote by $gcd\,(b_1, \ldots, b_n)$ the greatest common divisor of integers $b_1, \ldots, b_n$.

Lemma 2. *For any pairwise different $j_1, \ldots, j_k \in \{1, \ldots, T\}$*

$$\mathbf{P}\left(a_1 x_{j_1} + \ldots + a_k x_{j_k} = 0\right) = f(a_1, \ldots, a_k, G),$$

where

$$f(a_1, \ldots, a_k, G) = \frac{gcd\,(a_1, \ldots, a_k, |G_1|) \cdot \ldots \cdot gcd\,(a_1, \ldots, a_k, |G_l|)}{|G_1| \cdot \ldots \cdot |G_l|}$$

and $G = G_1 \times \ldots \times G_l$ is the representation of G as a product of primary cyclic groups.

Denote by $\mathcal{J}(k,T)$ the set of ordered tuples $(j_1,\dots,j_k)$, each of which consists of k pairwise different elements of the set $\{1,\dots,T\}$. Thus $|\mathcal{J}(k,T)| = \frac{T!}{(T-k)!}$. For any $(j_1,\dots,j_k) \in \mathcal{J}(k,T)$ consider the linear relation

$$a_1 x_{j_1} + \dots + a_k x_{j_k} = 0.$$

Theorem 1. *([3]) Let $a_1,\dots,a_k$ be fixed non-zero integers and let $x_1,\dots,x_T$ be independent random variables uniformly distributed on a finite additive Abelian group G and*

$$\zeta = \frac{1}{w(a_1,\dots,a_k)} \sum_{(j_1,\dots,j_k)\in\mathcal{J}(k,T)} I(a_1 x_{j_1} + \dots + a_k x_{j_k} = 0).$$

Let $G = G(T)$ be a sequence of groups which vary as $T \to \infty$ in such a way that $|G| \to \infty$ and

$$\mathbf{E}\zeta = \frac{T!}{(T-k)!} \cdot \frac{f(a_1,\dots,a_k,G)}{w(a_1,\dots,a_k)} \to \lambda,\ 0 < \lambda < \infty.$$

Assume further that for any fixed natural number b, any fixed $\varepsilon > 0$ and a random variable y uniformly distributed on the group G

$$\mathbf{P}(by = 0) = o\left(|G|^{-1+\varepsilon}\right).$$

Then

$$\lim_{|G|,T\to\infty} \mathbf{P}\left(\zeta = m\right) = \frac{\lambda^m}{m!} e^{-\lambda},\ m = 0,1,2,\dots\ .$$

The proof of the theorem is based on Sevastianov's method described in [2].

Remark.

- A sequence of groups $\mathbf{Z}_M^d$, where d is fixed and $M \to \infty$, meets the conditions of Theorem 1.
- If ζ is specified by the elements of a sequence of groups $\mathbf{Z}_q^d$, where q is a fixed prime number, then the distribution of ζ converges as $d \to \infty$ to a Poisson distribution as well. Observe that this convergence does not follow from Theorem 1.
- Let $G = \mathbf{Z}_4 \times (\mathbf{Z}_2)^{d-2}$ and

$$\zeta = \frac{1}{2} \sum_{(j_1,j_2,j_3)\in\mathcal{J}(3,T)} I(-x_{j_1} + x_{j_2} + x_{j_3} = 0).$$

  Assume $d \to \infty$ and $T \to \infty$ in such a way that

$$\mathbf{E}\zeta = \frac{3}{|G|} C_T^3 \to \lambda \in (0,+\infty).$$

  Then the distribution of ζ converges to the compound Poisson distribution.

2 A goodness-of-fit test

Theorem 1 can be used to construct a goodness-of-fit test to check the hypothesis that elements of a sample were generated by the uniform distribution. Let H_0 be the hypothesis that random variables $x_1, \ldots, x_T$ are independent and uniformly distributed on a fixed Abelian group G. Observe that if $a_1 = \ldots = a_k = 1$, then $w(a_1, \ldots, a_k) = k!$ and

$$\zeta = \frac{1}{k!} \sum_{(j_1,\ldots,j_k)\in\mathcal{J}(k,T)} I(x_{j_1} + \ldots + x_{j_k} = 0) = \sum_{1\le j_1<\ldots<j_k\le T} I(x_{j_1} + \ldots + x_{j_k} = 0).$$

Theorem 2. *If the hypotesis H_0 is true, then $\lambda = \mathbf{E}\zeta = \frac{C_T^k}{|G|}$ and*

$$d_{TV}(\zeta, \pi_\lambda) = \sup_{A\subset\mathbf{Z}_+} |\mathbf{P}(\zeta \in A) - \mathbf{P}(\pi_\lambda \in A)| \le k_2(\lambda)\,\lambda^2 \left(\frac{2k^2}{T-k+1} + \frac{1}{C_T^k}\right),$$

where π_λ is a Poisson random variable with parameter λ and $k_2(\lambda) = \left(1 - e^{-\lambda}\right)\lambda^{-1}$.

The proof of the theorem is based on the Chen-Stein method described in [1].
Criterion. *Let $\alpha_{\max} > 0$ be the maximal probability to reject H_0 if it is true and*

$$\alpha_m = \sum_{n=m+1}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!}, m = 0, 1, 2, \ldots .$$

Select $l = \min\{m : \alpha_m \le \alpha_{\max}\}$ and calculate $\zeta = \sum_{1\le j_1<\ldots<j_k\le T} I(x_{j_1} + \ldots + x_{j_k} = 0)$ for the sample $x_1, \ldots, x_T$. Hypothesis H_0 should be accepted if $\zeta \le l$ and rejected if $\zeta > l$.

This criterion allows to check H_0 on a group G by a sample of size having order $\sqrt[k]{|G|}$. It follows from Theorem 2 that the significance level α for H_0 admits the estimate

$$|\alpha - \alpha_l| \le k_2(\lambda)\,\lambda^2 \left(\frac{2k^2}{T-k+1} + \frac{1}{C_T^k}\right).$$

Example. Consider the hypothesis H_0 that $x_1, \ldots, x_{25000}$ is a sample generated by a random number generator, which generates independent elements uniformly distributed on group $\mathbf{Z}_{2^{40}}$. In this case $T = 25000$, and, if $k = 3$,

$$\zeta = \sum_{1\le j_1<j_2<j_3\le 25000} I(x_{j_1} + x_{j_2} + x_{j_3} = 0),$$

and $\lambda = \frac{C_{25000}^3}{2^{40}} \approx 2.36819$.

If $\alpha_{\max} = 0.05$, then $l = 5, \alpha_l \approx 0.03379$ and H_0 is accepted, if $\zeta \le l = 5$, and rejected, if $\zeta > l = 5$. The significance level α lies within the range 0.03379 ± 0.00154.

References

[1] Barbour A.D., Holst L., Janson S. (2002) *Poisson Approximation.* — Oxford University Press.

[2] Kolchin V.F., Sevastianov B.A., Chistiakov V.P. (1978) *Random allocations.* — V. H. Winston, New York.

[3] Kruglov V.I. (2008) *Limit distributions of the number of vectors satisfying a linear relation.* — (Russian) Diskretnaya Matematika, vol. 20, iss. 4, p. 120–135. Translation in Discrete Mathematics and Applications, vol. 18, iss. 5, p. 465–481.