Е. Л. Емельянова (Самара, СГАУ). **О группировке эмпирических данных с использованием равномерно распределенных последовательностей случайных чисел.**

Пусть наблюдается случайная величина X с неизвестной функцией распределения $F(x)$, $\overline{X} = (X_1, X_2, \ldots, X_n)$ — некоторая отвечающая ей случайная выборка и $\omega = (X_1, X_2, \ldots, X_n)$ — реализация этой выборки. Если объем выборки $\overline{X}$ велик, то, прежде чем исследовать эту выборку по существу, сначала разбивают вариационный ряд на части и производят осреднение элементов в каждой части. Естественно, эта процедура, привнося влияние посторонних факторов методического плана, ухудшает качество исходной выборки. Ниже предлагается методика группировки элементов выборки $\overline{X}$, уменьшающая влияние таких факторов.

Возьмем произвольно целое число $q \geqslant 2$ и равномерно распределенную в промежутке $[0,1)$ последовательность случайных чисел

$$\alpha_\ell = 0, \delta_\ell \delta_{\ell 1} \delta_{\ell+2} \ldots \qquad (\ell = 1, 2, \ldots), \tag{1}$$

записанных в q-ичной системе счисления, т. е. такую, что для любого промежутка $\Delta \subset [0,1)$ длины $|\Delta| < 1$ и любого натурального числа ℓ при $P \to \infty$ выполняется соотношение

$$N_P^{(\ell)}(\Delta) = |\Delta| P + o\,(P), \tag{2}$$

где $N_P^{(\ell)}(\Delta)$ — количество чисел $\alpha_{\ell-1+i}$, $i = 1, 2, \ldots, P$, попавших в промежуток Δ. При построении чисел (1) с $q = 10$ можно использовать, например, табл. 7 на с. 240 книги [1]; для решения стандартных практических задач этого вполне достаточно.

Взяв, далее, произвольно целое число $k \geqslant 2$, например, $k \in [n/h, n/h+1)$, где $h = (x_{\max} - x_{\min})/(1 + 3,322 \lg n)$ есть число Стэрджесса, для каждого $s = 1, 2, \ldots, k$ положим $\Delta_s = [(s-1)/k, s/k)$ и определим функцию $f_s(t)$, $0 \leqslant t < 1$, равенствами

$$f_s(t) = \chi_{\Delta_s}(t) = \begin{cases} 1 & \text{при } \{t\} \in \Delta_s, \\ 0 & \text{при } \{t\} \notin \Delta_s, \end{cases} \qquad 0 \leqslant t < 1. \tag{3}$$

Образуем k подвыборок $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$ выборки $\overline{X}$, относя к $\overline{X}_s$, $s = 1, 2, \ldots, k$, элементы X_i выборки $\overline{X}$ с номерами i, для которых $\alpha_{s-1+i} \in \Delta_s$. Очевидно, подвыборки $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$ попарно не пересекаются, а их объединение дает всю выборку $\overline{X}$. Разбиение выборки $\overline{X}$ на k частей $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$ естественным образом порождает и разбиение набора ω на части $\omega_1, \omega_2, \ldots, \omega_k$, являющиеся реализациями подвыборок $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$, соответственно. Пусть $F_n^*(x)$ — эмпирическая функция распределения случайной величины X, построенная по реализации ω случайной выборки $\overline{X}$, а при $s = 1, 2, \ldots, k$ функция $F_{n,s}^*(x)$ — эмпирическая функция распределения случайной величины X, построенная по реализации ω_s случайной подвыборки $\overline{X}_s$. Легко видеть, что

$$F_n^*(x) = \frac{1}{k} \sum_{s=1}^{k} F_{n,s}^*(x), \quad -\infty < x < \infty. \tag{4}$$

Если v_s — объем подвыборки $\overline{X}_s$, то в силу соотношений (3) и (2), при всех $s = 1, 2, \ldots, k$

$$v_s = \sum_{i=1}^{n} f_s(\alpha_{s-1+i}) = \frac{n}{k} + o\,(n),$$

когда $n \to \infty$, т.е. $v_s \approx [n/k]$, причем последнее приближенное равенство тем точнее, чем меньше модуль остаточного члена в соотношении (2).

В силу свойств последовательности (1), подвыборки $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$ можно рассматривать как почти независимые случайные выборки, отвечающие случайной величине X, причем качество независимости определяется скоростью стремления к нулю остаточного члена в соотношении (2). В связи с этим важную роль здесь могут сыграть построенные в теории чисел последовательности (в частности, последовательности дробных долей показательной функции с целым основанием $q \geqslant 2$) с быстрым равномерным распределением в промежутке $[0,1)$, ведущие себя как почти независимые случайные величины в соответствующем вероятностном пространстве (см. [2]–[5]).

Описанный метод разбиения случайной выборки $\overline{X}$ на части уменьшает влияние посторонних факторов методического плана (например, когда объединяются в группы соседние члены выборки $\overline{X}$) и, как следствие, повышает, в частности, надежность критериев согласия при проверке статистических гипотез, делая эти критерии менее чувствительными по отношению к процедуре группировки опытных данных. Поэтому, проверяя согласие непрерывной гипотетической функции распределения $F(x)$ с опытными данными, например, с помощью критерия А. Н. Колмогорова, целесообразно наряду с классической статистикой этого критерия

$$D_n(\overline{X}) = \sqrt{n} \sup_{-\infty<x<\infty} |F_n^*(x) - F(x)| \tag{5}$$

использовать также модифицированные статистики

$$D_n^{(k)}(\overline{X}) = \frac{1}{k} \sum_{s=1}^{k} \sqrt{\frac{n}{k}} \sup_{-\infty<x<\infty} |F_{n,s}^*(x) - F(x)|, \tag{6}$$

имеющие при каждом целом $k \geqslant 2$ то же самое предельное при $n \to \infty$ распределение, что и классическая статистика $D_n(\overline{X})$. Вследствие равенств (5) и (4),

$$D_n(\overline{X}) = \sqrt{n} \sup_{-\infty<x<\infty} \left| \frac{1}{k} \sum_{s=1}^{k} \Big(F_{n,s}^*(x) - F(x) \Big) \right|. \tag{7}$$

Поэтому, сравнивая соотношения (6) и (7), мы видим, что $D_n^{(k)}(\overline{X})$ $(k = 2, 3, \ldots)$ — это существенно другие статистики, нежели $D_n(\overline{X})$, хотя они и имеют такое же предельное при $n \to \infty$ распределение.

СПИСОК ЛИТЕРАТУРЫ

1. *Чистяков В. П.* Курс теории вероятностей. М.: Наука, 1982.
2. *Коробов Н. М.* О распределении дробных долей показательных функций. — Вестник МГУ, серия матем., мех., 1966, № 4, с. 42–46.
3. *Усольцев Л. П.* Задача на построение, связанная с равномерным распределением дробных долей показательных функций. — Известия ВУЗов, серия матем., 1967, № 12 (67), с. 75–83.
4. *Усольцев Л. П.*, *Емельянова Е. Л.* Модификация статистики Пирсона. — Обозрение прикл. и промышл. матем., 2005, т. 12, в. 2, с. 535–537.
5. *Емельянова Е. Л.* Теоретико-числовая модификация статистики А. Н. Колмогорова. — Обозрение прикл. и промышл. матем., 2005, т. 12, в. 4, с. 854–856.