УДК 519.21

С. А. МАТВЕЙЧУК, мл. науч. сотр.,
Ю. И. ПЕТУНИН, д-р физ.-мат. наук (Киев. ун-т)

Обобщение схемы Бернулли, возникающее в вариационной статистике. II

Изучается частота появления событий в модифицированной схеме Бернулли при предположении $F_x(u) \equiv F_y(u)$. Предлагается критерий однородности двух выборок, основанный на свойствах этой частоты и исследуются его свойства.

Вивчається частота появи подій у модифікованій схемі Бернуллі з припущенням, що $F_x(u) \equiv F_y(u)$. Пропонується критерій однорідності двох виборок, який грунтується на властивостях цієї частоти і досліджуються його властивості.

Введение. Пусть $\bar{x} = (x_1, \ldots, x_n)$ и $\bar{y} = (y_1, \ldots, y_m)$ — независимые выборки, полученные путем простого случайного выбора из генеральных совокупностей G_x и G_y с непрерывными строго возрастающими функциями распределения $F_x(u)$ и $F_y(u)$ соответственно. Построим по выборке $\bar{x}$ вариационный ряд $x^{(0)} < x^{(1)} \leqslant \ldots \leqslant x^{(n)} < x^{(n+1)}$ $(x^{(0)} = -\infty,\ x^{(n+1)} = \infty)$ и рассмотрим случайный интервал $\mathcal{I}_{i,q} = (x^{(i)}, x^{(i+q)}]$, где i, q — фиксированные числа $(1 \leqslant q \leqslant n,\ 0 \leqslant i \leqslant n - q + 1)$. Предположим, что $F_x(u) \equiv F_y(u)$. Схему испытаний, в которой на k-м шаге событие $A_k = \{y_k \in$

$\in \mathcal{J}_{i,q}\}$ (y_k — элемент выборки $\bar{y}$) происходит или не происходит, назовем модифицированной схемой Бернулли, построенной по выборкам $\bar{x}$ и $\bar{y}$. Пусть $\varkappa_{i,q}$ — число событий A_k, реализовавшихся в серии из m испытаний (m — объем выборки $\bar{y}$). Случайная величина $\varkappa_{i,q}$ изучалась в работе [1], где найдено ее распределение и исследованы ее асимптотические свойства при различных способах стремления n и m, объемов выборок $\bar{x}$ и $\bar{y}$ к бесконечности. Там же введено понятие обобщенной схемы Бернулли как модифицированной схемы Бернулли, в которой $F_x(u) \equiv F_y(u)$. Настоящая статья является продолжением [1]. Она посвящена изучению свойств статистики $\varkappa_{i,q}$ в обобщенной схеме Бернулли (п. 2), а так же построению критерия однородности выборок $\bar{x}$ и $\bar{y}$, основанного на этой статистике (п. 3).

Понятия и обозначения работы — такие же, как и в статье [1].

2. Обобщенная схема Бернулли. В этой схеме $F_x(u) \equiv F_y(u)$, поэтому идентификатор генеральных совокупностей [1] $G(v) = F_y[F_x^{-1}(v)]$, $v \in [0, 1]$, будет тождественно равен $v : G(v) \equiv v$, $v \in [0, 1]$. Учитывая этот факт, сформулируем такое определение обобщенной схемы Бернулли.

Определение 1. *Обобщенной схемой Бернулли называется такая модифицированная схема Бернулли, для которой идентификатор генеральных совокупностей есть единичное отображение отрезка* [0, 1] *в себя.*

На основании этого определения все утверждения, справедливые для модифицированной схемы Бернулли, непосредственно переносятся на обобщенную схему Бернулли.

Условимся в дальнейшем для случайной величины $\varkappa_{i,q}$ в обобщенной схеме Бернулли использовать обозначение $\chi_{i,q}$.

Из теоремы 1, § 1, [1] путем замены $G(v)$ на v и вычисления полученных интегралов можно получить следующее утверждение.

Утверждение 1. *В обобщенной схеме испытаний Бернулли случайная величина $\chi_{i,q}$ имеет следующее распределение вероятностей:*

$$\mathrm{P}(\chi_{i,q} = l) = \frac{q}{q+l} \frac{C_m^l C_n^q}{C_{m+n}^{l+q}}, \quad l = 0, 1, \ldots, m, \tag{1}$$

где n, m — объемы выборок, $\bar{x}$, $\bar{y}$ соответственно; i, q — фиксированные числа ($1 \leqslant q \leqslant n$, $0 \leqslant i \leqslant n-q+1$), определяющие интервал $\mathcal{J}_{i,q}$; C_m^n — число сочетаний из m по n.

Распределение (1) впервые получено Эпстейном для интервала $\mathcal{J}_{0,q} = (-\infty, x^{(q)}]$ [2].

Из (1) следует, что математическое ожидание $E(\chi_{i,q})$ и дисперсия $D(\chi_{i,q})$ случайной величины $\chi_{i,q}$ равны:

$$E(\chi_{i,q}) = mp_q, \tag{2}$$

$$D(\chi_{i,q}) = \frac{m(m+n+1)}{n+2} p_q(1-p_q), \tag{3}$$

где

$$p_q = \frac{q}{n+1}. \tag{4}$$

Будем изучать асимптотические свойства $\chi_{i,q}$. Пусть сначала n — объем выборки $\bar{x}$ — является фиксированной величиной, а m — объем выборки $\bar{y}$ — неограниченно возрастает. Вместо случайной величины $\chi_{i,q}$ будем рассматривать частоту $h_{i,q} = \chi_{i,q}/m$; тогда из теоремы 4, § 1, [1] вытекает такое утверждение.

Утверждение 2. *В обобщенной схеме испытаний Бернулли при фиксированном n и $m \to \infty$ частота $h_{i,q}$ сходится по распределению к непрерывной случайной величине h_n, имеющей следующую функцию распределения:*

$$\mathrm{P}(h_n \leqslant x) = \mathrm{I}_x(q, n-q+1), \quad x \in [0, 1], \tag{5}$$

где

$$\mathrm{I}_x(q,\ n-q+1)=B^{-1}(q,\ n-q+1)\int\limits_0^x t^{q-1}(1-t)^{n-q}\,dt \tag{6}$$

— *функция бета-распределения* [3], $B(q,\ n-q+1)=\int\limits_0^1 t^{q-1}(1-t)^{n-q}dt$ — *бета-функция Эйлера* [3].

Пусть теперь m фиксировано, а n неограниченно возрастает. Из теоремы 7, § 1, [1] путем замены $G(v)$ на v вытекает следующее утверждение.

Утверждение 3. *Если в обобщенной схеме Бернулли интервал $\mathcal{J}_{i,q}$ выбран таким образом, что при $n\to\infty$ $i/(n+1)\to p^*$, $q/(n+1)\to\to p_0\ (p^*,\ p_0\in(0,\ 1))$, то при фиксированном m и $n\to\infty$ случайная величина $\chi_{i,q}$ сходится по распределению к случайной величине ν_m, имеющей следующую функцию распределения:*

$$\mathrm{P}(\nu_m\leqslant k)=B_m(k,\ p_0),\qquad k=0,\ 1,\ \dots,\ m; \tag{7}$$

где

$$B_m(k,\ p_0)=\sum_{l=0}^{k}C_m^l p_0^l(1-p_0)^{m-l} \tag{8}$$

— *функция биномиального распределения* [4].

Рассмотрим теперь $\chi_{i,q}$, когда n и m стремятся к бесконечности одновременно. Введем нормированную случайную величину

$$\zeta_{i,q}=(\chi_{i,q}-E(\chi_{i,q}))/\sqrt{D(\chi_{i,q})}.$$

Из теоремы 10, § 1, [1] получим такое утверждение.

Утверждение 4. *В обобщенной схеме испытаний Бернулли при $n\to\infty$ и $m\to\infty$, если*

$$i/(n+1)\to p^*,\quad q/(n+1)\to p_0\quad (p^*,\ p_0\in(0,\ 1)),$$

то случайная величина $\zeta_{i,q}$ сходится по распределению к случайной величине ζ_0, имеющей стандартное нормальное распределение

$$\mathrm{P}(\zeta_0\leqslant x)=\Phi(x),\quad x\in R^1, \tag{9}$$

где

$$\Phi(x)=(\sqrt{2\pi})^{-1}\int\limits_{-\infty}^{x}e^{-y^2/2}dy. \tag{10}$$

Рассмотрим поведение $\chi_{i,q}$ при $m\to\infty$, $n\to\infty$, но интервал $\mathcal{J}_{i,q}$ при этом будем выбирать так, чтобы $i/(n+1)\to p^*$, а q было фиксированным числом.

Тогда из теоремы 13, § 1 [1] следует утверждение 5.

Утверждение 5. *В обобщенной схеме испытаний Бернулли при $m,\ n\to\infty$, если*

$$m/(n-2)\to r\ (r\in R^+),\qquad i/(n+1)\to p^*\quad (p^*\in(0,\ 1)),$$

q фиксировано, то случайная величина $\chi_{i,q}$ сходится по распределению к случайной величине η_q, имеющей следующую функцию распределения:

$$\mathrm{P}(\eta_q\leqslant k)=F_q(k)\equiv\sum_{l=0}^{k}C_{l+q-1}^{l}\cdot\frac{r^l}{(1+r)^{l+q}},\qquad k=0,\ 1,\ \dots; \tag{11}$$

где $F_q(k)$ — функция отрицательного биномиального распределения [4].

Этим исследование статистики $\chi_{i,q}$ обобщенной схемы Бернулли завершено.

3. Некоторые критерии однородности двух выборок. Используя свойства статистик $\varkappa_{i,q}$ и $\chi_{i,q}$, можно предложить ряд критериев однородности двух выборок. Рассмотрим некоторые из них.

Пусть $\bar{x}$ и $\bar{y}$ — две независимые выборки объема n и m соответственно, принадлежащие генеральным совокупностям G_x и G_y с функциями распределения $F_x(u)$ и $F_y(u)$. Будем считать, что $F_x(u)$ и $F_y(u)$ являются непрерывными, строго возрастающими функциями. Тогда существует идентификатор генеральных совокупностей $G(v) = F_y[F_x^{-1}(v)]$, который является непрерывно строго возрастающей функцией, отображающей отрезок $[0, 1]$ в себя. Обозначим через $\mathcal{G}[0, 1]$ класс всех таких функций $G(v)$. Проблема однородности двух выборок состоит в следующем: является ли идентификатор генеральных совокупностей $G(v)$ единичным отображением $\mathrm{I}[0, 1]$, отображающим отрезок $[0, 1]$ в себя; или же $G(v)$ принадлежит классу $\mathcal{G}[0, 1]\setminus \mathrm{I}[0, 1]$. Назовем первое предположение основной гипотезой $H_0 = \{G(v) \equiv v,\ v \in [0, 1]\}$, а второе предположение — сложной альтернативной гипотезой $H_1 = \{G(v) \in \mathcal{G}[0, 1]\setminus \mathrm{I}[0, 1]\}$.

Для того чтобы построить критерий или тест проверки гипотезы H_0 против альтернативы H_1, воспользуемся статистиками $\varkappa_{i,q}$ и $\chi_{i,q}$. В силу того что распределение $\chi_{i,q}$ не зависит от функций распределения выборок $F_x(u)$ и $F_y(u)$ (см. утверждения 1 — 5, 2), построим доверительный интервал $\mathcal{J}(\chi_{i,q}) = (\chi_{i,q}^{\text{н}}, \chi_{i,q}^{\text{в}})$, симметричный относительно математического ожидания $\chi_{i,q}$ и содержащий основную массу значений случайной величины $\chi_{i,q}$ с заданным уровнем значимости 2β:

$$\chi_{i,q}^{\text{н}} = E(\chi_{i,q}) - a\sqrt{D(\chi_{i,q})}, \tag{12}$$

$$\chi_{i,q}^{\text{в}} = E(\chi_{i,q}) + a\sqrt{D(\chi_{i,q})},$$

где $E(\chi_{i,q})$, $D(\chi_{i,q})$ — математическое ожидание и дисперсия случайной величины $\chi_{i,q}$ (см. (2) — (4)); величина a определяется из соотношения

$$\mathrm{P}(\chi_{i,q} \in \mathcal{J}(\chi_{i,q})) = 1 - 2\beta. \tag{13}$$

Предлагается следующий критерий проверки гипотезы H_0 против альтернативы H_1:

1) выбирается два натуральных числа q и i, где $1 \leqslant q \leqslant n$, $0 \leqslant i \leqslant n - q + 1$, n — объем выборки $\bar{x}$;

2) по заданному уровню значимости 2β и числам i, q, исходя из распределения (1) или одной из его аппроксимаций (утверждения 1—5), строится доверительный интервал $\mathcal{J}(\chi_{i,q})$ согласно формулам (12)—(13);

3) по выборке $\bar{x}$ строится вариационный ряд, а по числам i и q — случайный интервал $\mathcal{J}_{i,q} = (x^{(i)}, x^{(i+q)}]$;

4) вычисляется значение статистики $\theta_{i,q}$, равное числу элементов выборки $\bar{y}$, попавших в интервал $\mathcal{J}_{i,q}$;

5) если $\theta_{i,q} \in \mathcal{J}(\chi_{i,q})$, то принимается гипотеза H_0; в противном случае принимается гипотеза H_1.

Из описания предложенного критерия следует, что независимые выборки $\bar{x}$ и $\bar{y}$ образуют модифицированную схему Бернулли, если имеет место гипотеза H_1.

Статистика критерия $\theta_{i,q}$ при справедливости гипотезы H_1 совпадает со статистикой $\varkappa_{i,q}$, изученной в § 1 [1], а если верна гипотеза H_0 — со статистикой $\chi_{i,q}$, изученной в п. 2.

Критической областью критерия является интервал $\bar{\mathcal{J}}(\chi_{i,q}) = R^1 \setminus \mathcal{J}(\chi_{i,q})$, при этом $\mathbf{P}(\bar{\mathcal{J}}(\chi_{i,q})/H_0) = 2\beta$. Поэтому в дальнейшем описанный критерий будем именовать как тест $\bar{\mathcal{J}}(\chi_{i,q})$ размера 2β.

Исследуем свойства этого теста. Предположим, что имеет место гипотеза H_1. Обозначим $h_{i,q} = \theta_{i,q}/m$. Если учесть выражения для математического ожидания и дисперсии $\varkappa_{i,q}$ (см. (14)—(15) § 1 [1]), то из предложения 2 (см. введение [1]) следует

$$E(h_{i,q}/H_1) = p_{i,q} + O(n^{-1}), \tag{14}$$

$$D(h_{i,q}/H_1) = p_{i,q}(1 - p_{i,q})/m + O(n^{-1}), \tag{15}$$

где

$$p_{i,q} = G\left(\frac{i+q}{n+1}\right) - G\left(\frac{i}{n+1}\right). \tag{16}$$

Поскольку функция $G(v)$ непрерывна на отрезке $[0, 1]$, а $i/(n+1) \to p^*$, $q/(n+1) \to p_0$ (p^*, $p_0 \in (0, 1)$), когда $n \to \infty$, то

$$E(h_{i,q}/H_1) \to p_1 \equiv G(p^* + p_0) - G(p^*), \tag{17}$$

$$D(h_{i,q}/H_1) \to 0 \tag{18}$$

при $m, n \to \infty$.

Пусть теперь имеет место гипотеза H_0, тогда согласно результатам п. 2 математическое ожидание и дисперсия $\theta_{i,q}$ определяются выражениями (3) и (4) соответственно и при $m, n \to \infty$, когда $q/(n+1) \to p_0 \in (0, 1)$,

$$E(h_{i,q}/H_0) \to p_0, \tag{19}$$

$$D(h_{i,q}/H_0) \to 0, \tag{20}$$

где $h_{i,q} = \theta_{i,q}/m$.

Из соотношений (17)—(20) на основании неравенства Чебышева [5] следует такая теорема.

Т е о р е м а 1. *Если при* $m, n \to \infty$ $i/(n+1) \to p^* \in (0, 1)$, $q/(n+1) \to p_0 \in (0, 1)$, *то* 1) *в случае истинности гипотезы* H_1 *частота* $h_{i,q}$ *сходится по вероятности к* $p_1 = G(p^* + p_0) - G(p^*)$; 2) *в случае истинности гипотезы* H_0 *частота* $h_{i,q}$ *сходится по вероятности к* p_0.

Исследуем состоятельность теста $\overline{\mathcal{T}}(\chi_{i,q})$. Для этого покажем, что при $m, n \to \infty$

$$\mathbf{P}(\overline{\mathcal{T}}(\chi_{i,q})/H_1) \to 1.$$

Согласно (12) — (13)

$$\mathbf{P}(\overline{\mathcal{T}}(\chi_{i,q})/H_1) = \mathbf{P}\left(\varkappa_{i,q} \leqslant mp_q - a\sqrt{p_q(1-p_q)\,m(m+n+1)/(n+2)}\right) + \mathbf{P}\left(\varkappa_{i,q} \geqslant mp_q + a\sqrt{p_q(1-p_q)\,m(m+n+1)/(n+2)}\right).$$

На основании теоремы 9, § 1 [1] при $m, n \to \infty$ имеет место следующее асимптотическое равенство:

$$\mathbf{P}(\overline{\mathcal{T}}(\chi_{i,q})/H_1) = \Phi\left(\frac{p_q - p_{i,q}}{\sqrt{D(h_{i,q})}} - a\sqrt{\frac{(m+n+1)\,p_q(1-p_q)}{(n+2)\,mD(h_{i,q})}}\right) + \Phi\left(\frac{p_{i,q} - p_q}{\sqrt{D(h_{i,q})}} + a\sqrt{\frac{(m+n+1)\,p_q(1-p_q)}{(n+2)\,mD(h_{i,q})}}\right), \tag{21}$$

где $p_{i,q} = G\left(\frac{i+q}{n+1}\right) - G\left(\frac{i}{n+1}\right)$; $D(h_{i,q})$ — дисперсия случайной величины $h_{i,q}$ в случае истинности гипотезы H_1; $\Phi(u)$ — функция стандартного нормального распределения.

Если предположить, что $p_0 > p_1$ ($p_0 < p_1$) (см. теорему 1), то для достаточно больших m и n справедливо неравенство $p_q > p_{i,q}$ ($p_q < p_{i,q}$). Из него с учетом предельных соотношений (17) — (18) получаем, что $\mathbf{P}(\overline{\mathcal{T}}(\chi_{i,q})/H_1) \to 1$.

Таким образом, тест $\overline{\mathcal{T}}(\chi_{i,q})$ будет состоятелен при условии $p_0 > p_1$ ($p_0 < p_1$). Рассмотрим подробнее это условие. Согласно (17) $p_1 - p_0 = G(p^* + p_0) - G(p^*) - p_0$.

Поскольку $G(v) \in \mathcal{G}[0, 1]$, то почти всюду в $[0, 1]$ существует суммируемая конечная производная $G'(v) \equiv g(v)$, поэтому $p_1 - p_0 = \int_E (g(v) - 1)\,dv$, где $E \subset (p^*, p^* + p_0)$, $\mu E = p_0$ (μE означает Лебегову меру множества E). Для того чтобы интеграл, стоящий в правой части последнего равенства, был строго положителен (строго отрицателен), достаточно потребовать, чтобы почти всюду в E выполнялось неравенство $g(v) \geqslant 1$ ($g(v) \leqslant$

$\leqslant 1$), а множество точек E_1, на котором имеет место строгое неравенство, имело бы положительную меру.

Определим класс $\mathcal{G}[p^*, p^*+p_0] \subset \mathcal{G}[0, 1]$ функций $G(v)$ следующим образом: 1) почти всюду в интервале $(p^*, p^*+p_0)\, G'(v) \geqslant 1$ $(G'(v) \leqslant 1)$; 2) множество точек v таких, что $G'(v) > 1$ $(G'(v) < 1)$, $v \in (p^*, p^*+p_0)$, имеет положительную меру.

Рассмотрим гипотезу

$$H_1^* = \{G(v) \in \mathcal{G}[p^*, p^*+p_0]\},$$

тогда в силу изложенного выше справедлива теорема.

Теорема 2. *Если интервал $\mathcal{J}_{i,q}$ выбран таким образом, что при $n \to \infty$ $i/(n+1) \to p^*$, $q/(n+1) \to p_0$ $(p^*, p_0 \in (0, 1))$, то при $m, n \to \infty$ тест $\overline{\mathcal{J}}(\chi_{i,q})$ для проверки гипотезы H_0 будет состоятелен против альтернативы H_1^*.*

Из теоремы 2 следует, что при фиксированном интервале $\mathcal{J}_{i,q}$ в общем случае тест $\overline{\mathcal{J}}(\chi_{i,q})$ не будет состоятельным против альтернативы H_1. Однако, при фиксированном n существует $N = [n(n+1)/2]$ интервалов вида $\mathcal{J}_{i,q} = (x^{(i)}, x^{(i+q)}]$ ($[x]$ означает целую часть числа x).

Обозначим через $\mathcal{J}_N(\overline{x})$ класс всех интервалов $\mathcal{J}_{i,q} = (x^{(i)}, x^{(i+q)})$, которые можно построить с помощью любых двух порядковых статистик при фиксированном n:

$$\mathcal{J}_N(\overline{x}) = \{\mathcal{J}_{i,q} : \mathcal{J}_{i,q} = (x^{(i)}, x^{(i+q)}], \quad 1 \leqslant q \leqslant n,\ 0 \leqslant i \leqslant n-q+1\}.$$

Справедлива такая теорема.

Теорема 3. *В классе $\mathcal{J}_N(\overline{x})$ существует такой интервал $\mathcal{J}_{i^*,q^*}$, что при условиях $i^*/(n+1) \to p^*$, $q^*/(n+1) \to p_0$ $(p^*, p_0 \in (0, 1))$, когда $n \to \infty$, тест $\overline{\mathcal{J}}(\chi_{i^*,q^*})$ гипотезы H_0 будет состоятельным против альтернативы H_1 при $m, n \to \infty$.*

Доказательство. Предположим, что имеет место гипотеза H_1. Тогда $G(v) \not\equiv v$ и $G(v)$ — непрерывная строго возрастающая функция. Рассмотрим функцию $H(v) = G(v) - v$, которая является непрерывной с конечным изменением. Следовательно, почти всюду в $[0, 1]$ существует конечная производная $H'(v)$, которая является суммируемой функцией. Выберем точку $v_0 \in [0, 1]$ такую, что $H'(v_0) \neq 0$. Существование указанной точки v_0 следует из того, что $G(v) \not\equiv v$; но тогда на основании известной теоремы анализа (см., например, [6, с. 223]) получаем, что существует окрестность точки

$$V_{0,\varepsilon} = (v_0 - \varepsilon,\ v_0 + \varepsilon) \subset [0, 1],$$

в которой функция $H(v)$ строго монотонна. Среди всех точек v_0, в которых $H'(v) \neq 0$, выберем точку v_0^*, имеющую наибольшую связную открытую окрестность $V_{0,\varepsilon^*} = (v_0^* - \varepsilon^*, v_0^* + \varepsilon^*) \subset [0, 1]$, в которой функция $H(v)$ монотонна.

Зафиксируем ε^* и точку v_0^*. Положим $q^* = [2\varepsilon^*(n+1)]$, $i^* = [(v_0^* - \varepsilon^*)(n+1)]$ и построим интервал $\mathcal{J}_{i^*,q^*} = [(x^{(i^*)}, x^{(i^*+q^*)}]$. В силу выбора ε^* и v_0^* справедливы неравенства $1 \leqslant q^* \leqslant n$, $0 \leqslant i^* \leqslant n - q^* + 1$. Следовательно, $\mathcal{J}_{i^*,q^*} \in \mathcal{J}_N(\overline{x})$. Построим теперь тест $\overline{\mathcal{J}}(\chi_{i^*,q^*})$. Поскольку при $n \to \infty$ $i^*/(n+1) \to v_0^* - \varepsilon^*$, $q^*/(n+1) \to 2\varepsilon^*$, а функция $H(v)$ строго монотонна в интервале $(v_0^* - \varepsilon^*, v_0^* + \varepsilon^*)$, то $G(v)$ будет принадлежать классу $\mathcal{G}[v_0^* - \varepsilon^*, v_0^* + \varepsilon^*]$, и в силу теоремы 2 при $m, n \to \infty$ $\mathbf{P}(\overline{\mathcal{J}}(\chi_{i^*,q^*})/H_1) \to 1$. Теорема доказана.

4. Практические рекомендации при построении критериев. Рассмотрим следующую модификацию проблемы проверки гипотез о равенстве гипотетических функций распределения на основании обучающих выборок, которая достаточно часто встречается в приложениях. Пусть G_1 и G_2 — две генеральные совокупности, заданные своими обучающими выборками $\overline{x}_i = (x_{i1}, \ldots, x_{in_i}) \in G_i$ $(i=1, 2)$, а их гипотети-

ческие функции распределения $F_i(u)$ $(i = 1, 2)$ считаются неизвестными. Пусть, далее, $\bar{x} = (x_1, \ldots, x_n)$ — выборка, полученная путем простого случайного выбора из какой-либо одной генеральной совокупности G_i $(i=1, 2)$; требуется определить номер этой генеральной совокупности G_i (т. е. установить генеральную совокупность, из которой взята выборка $\bar{x}$). Эту задачу можно решить с помощью классического порядкового критерия Колмогорова — Смирнова путем вычисления статистики (или метрики)

$$\rho(F^*(u), F_i^*(u)) = \sup_{u \in R^1} |F^*(u) - F_i^*(u)|, \tag{22}$$

где $F^*(u)$, $F_i^*(u)$ — эмпирические функции распределения, полученные на основании выборок $\bar{x}$ и $\bar{x}_i$ соответственно, и построения доверительных интервалов для этой статистики, отвечающих заданному уровню значимости. Однако эта классическая метрика, наряду со значительными достоинствами, обладает рядом недостатков, из которых укажем здесь лишь один, связанный с видом метрики (22). Легко видеть, что эта метрика является обычной чебышевской метрикой в классе ограниченных функций со счетным множеством точек разрыва. Чебышевская метрика является достаточно грубой, поскольку она не в состоянии учесть различие производных функций распределения, т. е. их плотностей вероятности. В приложениях, однако, это различие возникает довольно часто. Например, в онкоморфологии, при дифференциальной диагностике опухолей, изучается модель генеральных совокупностей G_i, когда $F_2(u) = (1 - \alpha) F_1(u) + \alpha\Phi(u)$, где $\alpha \in (0, 1)$, а $\Phi(u)$ — некоторая функция распределения, плотность вероятности которой сосредоточена на достаточно малом носителе $[\alpha - \varepsilon, \alpha + \varepsilon]$. Ясно, что с помощью метрики (22) при соответствующем выборе α и ε мы не можем заметить различия между распределениями $F_1(u)$ и $F_2(u)$; это можно делать лишь с помощью более тонких дифференциальных метрик, в которых используются плотности вероятностей $f_1(u) \equiv F_1'(u)$ и $f_2(u) \equiv F_2'(u)$. Можно показать, что введенные в этой работе статистики $\varkappa_{i,q}$ и $\chi_{i,q}$ как раз и являются мерами близкости между выборками $\bar{x}$, $\bar{x}_i$, которые относятся к типу более тонких, чем грубая метрика (22), дифференциальных мер близости. В связи с этим построенные выше критерии рекомендуем использовать при идентификации функций распределения в моделях смесей распределений

$$F_2(u) = (1 - \alpha) F_1(u) + \alpha\Phi(u).$$

Перейдем теперь к проблеме построения интервала $\mathcal{I}_{i,q}$, на котором основаны предложенные в этой работе критерии. Ясно, что этот интервал следует выбирать в той области (a, b) числовой прямой, где наиболее велико различие (например, в метрике C или $\mathcal{L}_1$) плотностей вероятности $f_1(u)$ и $f_2(u)$. Поскольку эти плотности вероятностей являются неизвестными, то их следует заменить на гистограммы $f_1^*(u)$ и $f_2^*(u)$, построенные с помощью известных методов оценки гистограмм по обучающим выборкам $\bar{x}_1$, $\bar{x}_2$ соответственно. Определить интервал (a, b), где наблюдается наибольшее различие $f_1^*(u)$ и $f_2^*(u)$, с помощью графиков этих гистограмм обычно не трудно; зная этот интервал (a, b), можно найти порядковые статистики $x_1^{(i)}$ и $x_1^{(i+q)}$, построенные по выборке $\bar{x}_1$, которые содержат интервал (a, b) или образуют интервал $\mathcal{I}_{i,q}$, мало отличающийся от интервала (a, b).

1. *Матвейчук С. А., Петунин Ю. И.* Обобщение схемы Бернулли, возникающее в вариационной статистике. I // Укр. мат. журн.— 1990.— 42, № 4.— С. 518—528.
2. *B. Epstein.* Tables for distribution of the number of exceedances // Ann. Math. Stat.— 1954.— 25.— P. 762—768.
3. *Справочник* по специальным функциям / Под ред. Абрамовица М., Стиган И.— М. : Наука, 1979.— 832 с.
4. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики.— М. : Наука, 1983.— 416 с.

5. *Крамер Г.* Математические методы статистики.— М. : Мир, 1975.— 648 с.
6. *Фихтенгольц Г. М.* Курс дифференциального и интегрального исчисления : В 3-х т.— М. : Наука, 1966.— Т. 1.— 607 с.

Получено 07.08.89