Е.В. Черепанов
(*Москва*)

СТОХАСТИЧЕСКОЕ ОПИСАНИЕ ВЫБОРОЧНОГО МЕТОДА

В статье рассмотрена связь между случайным и квотным отборами при использовании выборочного метода в социологии. Найдены строгие соотношения для распределений, которым подчиняется случайный отбор. Показано, что распределение квотного отбора является допустимой аппроксимацией случайного, поскольку математические ожидания этих распределений совпадают. Предложен метод статистического оценивания частот встречаемости признаков (свойств) по случайным выборкам.

Ключевые слова: структурированная генеральная совокупность, репрезентативный отбор, стохастическое описание, статистические оценки, гипергеометрическое распределение, дихотомические переменные, случайная выборка, квотная выборка.

Один из крупнейших статистиков XX в. Дж.У. Тьюки писал: «К вероятности (в приложениях. – *Авт*.) следует относиться серьезно или оставлять ее в покое, если время от времени это может оказаться необходимым или полезным» [1]. В социологии невозможно даже на время оставить понятие случайности «в покое». Это обусловлено тем, что социальные феномены имеют настолько сложный характер, что проявляются только для значительных по численности групп населения. Следовательно, *социальные системы имеют вероятностную природу*, а их «экспериментальное» изучение возможно только методами, основанными на асимптотических свойствах стохастических систем.

Евгений Васильевич Черепанов – кандидат технических наук, зав. кафедрой социально-экономического и политического менеджмента Академии менеджмента инноваций. E-mail: e.cherep@gmail.com.

Важно подчеркнуть, что *случайность* и *вероятность* – базовые, понимаемые на интуитивном уровне понятия (как множество, точка, линия нулевой толщины и т.п.). Такие понятия нельзя определить, их можно только описать. Например, «вероятность» логически завершенно может быть описана статистически по Р. Мизесу [2] или на основе теоретико-множественной аксиоматики А.Н. Колмогорова [3]. Показательно, что в литературе (и даже учебного характера) можно встретить выражение «*определение вероятности* по Колмогорову». Это некорректно, ибо речь идет *об описании вероятности* А.Н. Колмогоровым, что впервые было им сделано в работе [4] с помощью модели «вероятностного пространства» (абстрактного множества с определенными на нем сигма-алгеброй и мерой, нормированной на единицу). Причем вопрос о границах применимости стохастического формализма за пределами естественнонаучных приложений специфичен и непрост [5; 6].

С момента возникновения прикладной социологии в 30-х гг. прошлого века ее методологическую основу составляет выборочный метод [7]. Мысль об его использовании базируется на законе больших чисел (в форме Я. Бернулли) [8], согласно которому выборочная частота встречаемости изучаемого признака в серии независимых опытов асимптотически (по объему выборки) сходится к истинной частоте. Но возникает принципиальная сложность: кроме условий случайности наблюдений и их независимости, теорема Я. Бернулли требует их априорной однородности [9]. Но любой социум – *структурированное*, причем по многим переменным, множество. При относительно небольших объемах выборки (скажем, 2-3 тыс. *случайно опрошенных* респондентов) различия в структурах генеральной совокупности и случайной выборки могут резко испортить точность оценки частоты встречаемости *исследуемого признака*.

Существует лишь два способа выхода из этой ситуации: 1) при расчетах математически строго учесть различия в структурах

случайного выборочного ансамбля и генеральной совокупности и 2) построить *квотную выборку,* которая по основным характеристикам (пол, возраст, образование, национальность и т.п.) была бы *репрезентативна* генеральной совокупности. Поскольку в 30-х гг. XX в. вычислительной техники не существовало, то у пионеров прикладной социологии, в общем-то, и выбора не было: раз считать условные вероятности не на чем, будем создавать квотные выборки. Попутно заметим, что создание квотной выборки для большой территории – занятие непростое, трудоемкое и дорогостоящее, как показано, например, М.С. Косолаповым [10].

Специфика социальных исследований состоит в том, что приходится работать преимущественно с нечисловыми данными, т.е. с переменными, выраженными в слабых шкалах [5; 11], применительно к которым само понятие «измерение» [12; 13] по смыслу отличается от измерений в сильных шкалах [14]. Основными типами слабых шкал являются порядковые и номинальные шкалы (или шкалы наименований) [11]. Особое место в социологических исследованиях занимает «предельный случай» шкалы наименований – булевы (или дихотомические) шкалы [11], порождаемые бинарными отношениями на множествах [16].

Статистический анализ нечисловых данных является частью такой сложной аналитической задачи, как обработка разнотипных переменных. В разных видах анализа она решается специфическим образом, например, так, как это преложено в типологическом анализе [15]. Вместе с тем, существует два базовых подхода вне зависимости от видов анализа.

Во-первых, «оцифровка» слабых переменных [17; 18]. Но объективно обосновать усиление шкалы измерения трудно, а тип «оцифровки» существенно предопределяет итоговые результаты. Причем любое усиление шкалы измерения является прямым «домысливанием» за природу каких-то дополнительных свойств исследуемой социальной системы. Даже традиционная замена порядковой шкалы [19] ранговой, позволяющая проводить серьез-

ный анализ эмпирических данных [20; 21], не может рассматриваться как универсальный метод. Поскольку конечное множество объектов социальной природы с заданным на нем отношением строгого порядка хотя и эквивалентно [16] некоторому подмножеству натурального ряда, по своей сути остается нечисловым множеством.

Во-вторых, подход, основанный на «дихотомизации», т.е. *на ослаблении всех переменных до булевого уровня с соответствующим увеличением размерности пространства признаков.* Идея подхода, который базируется на анализе статистик бинарного отношения на множествах [22], состоит в том, что сложный объект можно с примерно равной информативностью описать или небольшим числом сильных переменных, или большим числом слабых. Эта мысль близка взглядам О. Курно, которые он полтора века назад изложил в классическом трактате [23]. По О. Курно, любое сложное свойство объекта может быть представлено как суперпозиция его более простых свойств. Каждое их этих «более простых» свойств является комбинацией «еще более простых» и т.д. Иначе говоря, имеется возможность декомпозиции нечисловых свойств объекта до некоторого «элементарного» уровня. В итоге мы получаем набор дихотомических (или булевых, или бинарных) переменных, описывающих, с удовлетворяющей нас точностью, изучаемую стохастическую систему. При составлении любого вопросника следуют путем, указанным О. Курно, реализуя *принцип дихотомизации* формального описания исследуемого социума. Изучение аспектов интересующих проблем доводится до того «элементарного» уровня описания, который считается достаточным для практических выводов. Каждый вариант ответа на вопросы анкеты представляет собою булеву переменную («да» – в данном наблюдении зафиксирован дихотомический признак, «нет» – зафиксировано его отсутствии). В итоге опрос определяет бинарное отношение на множествах «респонденты» – «признаки» (варианты ответов).

В этой связи видится полезным как для социологов, так и для математиков, связанных с компьютерной обработкой данных, формально строго рассмотреть вероятностную сторону выборочного метода как основы эмпирических социологических исследований. И главное, требует ответа вопрос о том, *насколько правомерно использование вероятностного формализма* на квотных выборках. Они хотя и отражают многомерную структуру генеральной совокупности, но по самому своему построению являются *не вполне случайными* выборками. Поиску ответа на этот вопрос и посвящена статья. Кроме того, в работе дан подход к выборочному оценивания частот появления нечисловых признаков при работе со случайными выборками, который позволяет *значительно* повысить точность результатов по сравнению с традиционными методами квотного оценивания. Отметим, что выводы о характеристиках распределений случайного отбора ниже приводятся без доказательств, их можно найти в монографии [24].

Гипергеометрическое распределение (ГГР) вероятностей

Статистические процедуры, на которых базируется выборочный метод в социологических исследованиях, основаны на ГГР [25], что впервые в отечественных изданиях, насколько это известно автору, было отмечено в переводе книги У. Кокрена [7].

Пусть задана генеральная совокупность, представляющая собою население (избирателей, покупателей и т.д.), состоящая из N человек ($N >> 1$). Среди населения существует M человек, обладающих интересующим нас признаком (состоят в данной партии, собираются купить автомашину, являются клиентами пенсионного фонда, пользуются страховой услугой). Производится *случайная выборка* респондентов объема n. Вероятность того события, что в выборку попадут *ровно* m лиц, обладающих изучаемым дихотомическим признаком ($0 \le m \le n$), дается формулой:

$$\Pr\{m \mid n\} \equiv hy(m \mid M, N; n) = \binom{N}{n}^{-1} \binom{M}{m} \binom{N-M}{n-m}, \qquad (1)$$

где число $\binom{N}{n} \equiv \frac{N!}{(N-n)!\, n!}$, $\Pr\{\ldots\}$ обозначает вероятность события $\{\ldots\}$, а $hy(\ldots)$ – стандартное обозначение ГГР [26]. Математическое ожидание ГГР равно

$$\mu = n\nu;\ \nu = M / N,\ 0 \le \nu \le 1, \qquad (2)$$

а его дисперсия определяется как $Dm = n\nu\,(1-\nu)\frac{N-n}{N-1}$. (3)

Функция распределения ГГР определена в виде:

$$Hy(m \mid M, N; n) = \binom{N}{n}^{-1} \sum_{k=0}^{m} \binom{M}{k} \binom{N-M}{n-k} \qquad (4)$$

Отметим, что ГГР возникло из задач анализа качества массовой продукции [26]. Но сегодня многомерные обобщения ГГР могут быть применяться для корректного описания многих задач в социологии и маркетинге потребительских рынков, в банковском деле, при подготовке рекламных и избирательных кампаний, для обоснования проектов в лотерейном бизнесе и для актуарных расчетов в страховом деле.

Приведем два простейших примера.

Пример 1. Пусть среди населения города с численностью N страховой компанией было застраховано n жителей. Стохастически устойчивая вероятность наступления страхового случая за заданный период времени равна $\nu = M / N$, $(0 \le \nu << 1)$. Какова вероятность, что за это время страховой компании придется выплатить деньги по ровно m страховкам? Ответ дается ГГР вероятностей, задаваемым соотношением (1).

Пример 2. Пусть тираж «моментальной» лотереи равен N. В городе распространено n лотерейных билетов. Вероятность выигрыша равна $\nu = M / N$, $(0 \le \nu << 1)$. Какова вероятность того, что

в городе куплено ровно m выигрышных билетов? И здесь, очевидно, ответ дается ГГР вероятностей, определенным соотношением (1).

Используя понятие гамма-функции («Г-функции») можно получить удобное (для программной реализации) выражение для вычисления значений ГГР. Для положительных действительных $x \in \Re^{\oplus}$ Г-функция определена интегралом Эйлера II рода [27; 28]:

$$\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt = \int_0^1 (-\ln t)^{x-1} dt;\ x \in \Re^{\oplus}. \qquad (5)$$

На множестве комплексных чисел $z \in C$ Г-функция представима контурным интегралом Ганкеля [27; 28] в виде

$$\Gamma(z) = 2\pi\sqrt{-1} \Big/ \int_{-\infty}^{0} e^{t} t^{-z} dt;\ -\pi \le \arg t \le \pi;\ z \in C. \qquad (6)$$

Выражение (6) дает прямое аналитическое продолжение функции (5) в комплексную область. Особенности функции $\Gamma(z)$ вида наблюдаются в точках $z = -1, -2, \ldots$, где $\Gamma(z)$ имеет простые полюсы.

Для нас важно, что для всех натуральных чисел k выполняется: $\Gamma(k+1) = k!$, причем $\Gamma(0) = 1 = \Gamma(1)$. Это позволяет представить ГГР в виде:

$$hy(m \mid N, M; n) = \left[\Gamma(N - M + m - n + 1)\right]^{-1} \times$$
$$\times \frac{\Gamma(n+1)}{\Gamma(m+1)} \cdot \frac{\Gamma(M+1)}{\Gamma(N+1)} \cdot \frac{\Gamma(N-n+1)}{\Gamma(n-m+1)} \cdot \frac{\Gamma(N-M+1)}{\Gamma(M-m+1)}. \qquad (7)$$

Используя свойство Г-функции вида (см. [27]) $\Gamma(z-k) = \Gamma(z)\prod_{l=1}^{k}(z-l)^{-1}$, из (7) несложно получить: $hy(m \mid N, M; n) =$

$$= \left(1 + \frac{m}{N-n+1}\right)\left(\prod_{l=1}^{M}\left(1 - \frac{n-m}{N-l+1}\right)\right)\left(\prod_{l=1}^{m}\frac{(M-l+1)(n-l+1)}{l(N-n+1+l)}\right) \qquad (8)$$

Это выражение действительно «выгодно» отличается от традиционных представлений ГГР (в смысле его использования для машинных расчетов), которые основаны на приближенных вычислениях бесконечных сумм или произведений.

Используя выражение (3), можно получить простые оценки для гарантированной точности прямых оценок $\hat{\nu} = m/n$ частот встречаемости признаков. Непараметрическое правило «трех сигм» [9; 29], оценки «сверху» для погрешности частоты встречаемости имеют вид

$$\delta \le 3\sqrt{D\hat{\nu}} \cong 3\sqrt{\hat{\nu}(1-\hat{\nu})/n} \le \frac{3}{2\sqrt{n}}. \qquad (9)$$

Из этого соотношения легко получить гарантированную погрешность оценки частоты встречаемости признака в зависимости от n. Интересно и обратное: каковы должны быть объемы выборки для заданных уровней гарантированной погрешности. Преобразуя неравенство (9), получаем: $n \ge \frac{9}{4\delta^2}$. Соответствующие данные приведены в табл. 1.

Таблица 1

НЕОБХОДИМЫЕ ОБЪЕМЫ ВЫБОРКИ ДЛЯ ЗАДАННЫХ УРОВНЕЙ ГАРАНТИРОВАННОЙ ПОГРЕШНОСТИ

$\delta \le$	0,01	0,02	0,03	0,04	0,05	0,10
$n \ge$	22 500	5 600	2 500	1 400	600	225

Заметим, что для традиционных в социологии объемов выборки порядка 1,5-2 тыс. наблюдений гарантированная погрешность частоты равна примерно 3,5%.

Важно также отметить, что если нас интересуют статистические выводы по некоторой немногочисленной категории населения, то численность этой категории в репрезентативной выборке должна составлять (при минимально разумном пороге точности в 5%) не менее 600 (!) человек. Даже для обеспечения погрешности

в 10% нам потребуется 225 лиц из указанной немногочисленной категории. Практически это нереализуемо, следовательно, для анализа структуры общественного мнения следует организовывать отдельные исследования для каждой такой категории населения.

Многомерное гипергеометрическое распределение (МГГР)

Пусть изучается отношение населения региона, число жителей которого равно N, к некоторому вопросу, позицию по которому мы свели к r точкам зрения, причем каждой из этих точек зрения придерживаются $M_1, M_2, \ldots, M_r \left(\sum_j^r M_j \equiv M \le N \right)$ жителей. Тем самым, определяется номинальная шкала, порождающая разбиение жителей на $r + 1$ непересекающееся подмножество. Последнее из них, мощностью $N - M$, состоит из лиц, не имеющих точки зрения по изучаемой проблеме.

Допустим, что среди населения проведен *случайный* опрос n жителей, в результате чего в выборке зафиксировано $m_1, m_2, \ldots, m_r \left(\sum_j^r m_j \equiv m \le n \right)$ лиц, придерживающихся каждой из выделенных точек зрения. Какова вероятность того, что будут зафиксированы именно эти значения $m_1, m_2, \ldots, m_r$? Ответ на вопрос дается распределением, которое в дальнейшем будем называть МГГР.

Его аналитическое выражение легко получается из следующих соображений. Видим, что для двумерного случая

$$hy2(m_1, m - m_1 \mid M_1, M - M_1, N; n) = hy(m_1 \mid M_1, N; n).$$

Откуда, используя рекурсивное выражение вида:

$$\Pr(m_1, m_2, \ldots, m_k \mid n) = \Pr(m_k \mid m_1, \ldots, m_{k-1}) \Pr(m_1, \ldots, m_{k-1} \mid n),$$

несложно получить выражение для *r-мерного гипергеометрическим распределением (МГГР) вероятностей вида*:

$$hyr(\vec{m} \mid \vec{M}, N; n) = \binom{N}{n}^{-1} \binom{N - \sum_{j}^{r} M_j}{n - \sum_{j}^{r} m_j} \prod_{j}^{r} \binom{M_j}{m_j}; \vec{m}, \vec{M} \in \Re_r^{\oplus}. \quad (10)$$

Введя априорные частоты: $\nu_j \equiv M_j / N;\ j = \overline{1,r},$ (11)
выражение (10) можно представить в виде

$$hyr(\vec{m} \mid \vec{\nu}, N; n) = \binom{N}{n}^{-1} \binom{N(1 - \sum_{j}^{r} \nu_j)}{n - \sum_{j}^{r} m_j} \prod_{j}^{r} \binom{M\nu_j}{m_j}; \vec{m}, \vec{\nu} \in \Re_r^{\oplus}. \quad (12)$$

Значения математических ожиданий компонент случайного вектора имеют вид:

$$\mu_j = \mu(m_j \mid N, \vec{\nu}; n) = n M_j / N = n\nu_j;\ j = \overline{1,r}. \quad (13)$$

Значения их дисперсий определены как

$$\sigma_j^2 = D(m_j \mid N, \vec{\nu}; n) = \frac{n M_j}{N^2} \frac{(N - M_j)(N - n)}{N - 1} =$$

$$= n\nu_j(1 - \nu_j)\frac{N - n}{N - 1} \quad j = \overline{1,r}. \quad (14)$$

Ковариации компонент стохастического вектора, подчиненного МГГР, запишутся в виде

$$C_{jk} \equiv Cov(m_j, m_k \mid \nu_j, \nu_k, N; n) = -n\nu_k\nu_j \frac{N - n}{N - 1};\ j \neq k. \quad (15)$$

Для социологических исследований значение МГГР состоит в следующем. Пусть в инструментарии опроса содержится вопрос, определяющий некоторую номинальную шкалу. Например, «*За кого из кандидатов Вы намерены голосовать на предстоящих выборах Президента РФ?*». Варианты ответов имеют вид: 1) За Г.А. Зюганова; 2) За И.М. Хакамаду; … r) За В.В. Жири-

новского. За каждого из кандидатов (допустим, что нам это известно) намерены голосовать $N\nu_k$, $(k=\overline{1,r})$ человек, где N – число избирателей. Пусть при случайном отборе n респондентов из числа избирателей мнения соответственно нумерации распределились как $m_1, m_2, \ldots, m_r$, а $n-\sum_k^r m_k \geq 0$ респондентов не определились в выборе. Вероятность этого события вычисляется выражением (12).

Многомерное структурированное ГГР (МСГГР)

Пусть на множестве населения определено разбиение на r непересекающихся подмножеств, мощности которых обозначим в виде $N_1, N_2, \ldots, N_r; \left(\sum_j^r N_j = N\right)$, заданное некоторой шкалой наименований. В качестве такой номинальной шкалы могут выступать признаки, фиксируемые Госкомстатом: «пол», «возраст», «район проживания», «уровень образования», «условия проживания» и т.п. Важно, что для генеральной совокупности априори известны численности соответствующих групп населения $N_1, N_2, \ldots, N_r$ (или, что эквивалентно, известны частоты $\theta_1, \theta_2, ..., \theta_r$):

$$\theta_j = N_j / N,\ j=\overline{1,r};\ \vec{\theta} \equiv \{\theta_1, \theta_2, ..., \theta_r\};\ \left(\sum_j^r \theta_j \equiv 1\right) \qquad (16)$$

Предположим, что изучается частота встречаемости *некоторого признака* (намерение поддержать «Единую Россию» на предстоящих выборах; желание купить «Форд-Фокус»; намерение воспользоваться услугами страховой компании «МАКС» и т.п.). Пусть нам известно, что среди выделенных категорий населения этим признаком обладают ровно $M_1, M_2, \ldots, M_r; \left(\sum_j^r M_j \equiv M \leq N\right)$ граждан.

Определим вектор частот $\eta_1, \eta_2, ..., \eta_r$ в виде:

$$\eta_j = M_j / N_j = \nu_j / \theta_j, \quad j = \overline{1,r}; \quad \vec{\nu} \equiv \{\nu_1, \nu_2, ..., \nu_r\} \in \Re_r^{\oplus}. \quad (17)$$

Эти частоты соответствуют долям лиц, обладающих данным признаком, «внутри» каждой из выделенных категорий. Пусть проведен *случайный* опрос. Какова вероятность того *составного* события, что в случайную выборку объема n ($n << N$) попадет ровно $n_1, n_2, \ldots, n_r$ представителей выделенных групп населения, причем ровно $m_1, m_2, \ldots, m_r$ граждан каждой группы будут обладать изучаемым признаком? Несложно понять, что правомерно соотношение

$$\Pr\{\vec{m}, \vec{n} \mid \vec{M}, \vec{N}, n\} = \Pr\{\vec{n} \mid \vec{N}\} \prod_{j=1}^{r} \Pr\{m_j \mid n_j\}. \quad (18)$$

Отсюда следует выражение для искомой вероятности, которое будем называть *многомерным структурированным гипергеометрическим распределением* (МСГГР):

$$\hbar yr(\vec{n}, \vec{m} \mid \vec{\nu}, \vec{\theta}, N; n) = \binom{N}{n}^{-1} \prod_j^r \binom{M_j}{m_j} \binom{N_j - M_j}{n_j - m_j}. \quad (19)$$

Выражение для МСГГР можно также представить в виде

$$\hbar yr(\vec{n}, \vec{m} \mid \vec{\nu}, \vec{\theta}, N; n) = \binom{N}{n}^{-1} \prod_j^r \binom{N\nu_j}{m_j} \binom{N(\theta_j - \nu_j)}{n_j - m_j}. \quad (20)$$

Условное математическое ожидание каждой из компонент вектора $\vec{m} = (m_1, m_k, \ldots, m_r)$ определятся в виде:

$$\overline{m}_k(\vec{n}) \equiv M(m_k \mid \vec{n}) = \sum_{m_1=0}^{n_1} \sum_{m_2=0}^{n_2} \ldots \sum_{m_r=0}^{n_r} m_k \, \hbar yr(\vec{n}, \vec{m} \mid \vec{\nu}, \vec{\theta}, N; n) =$$

$$= \frac{n_k \nu_k}{\theta_k} \binom{N}{n}^{-1} \prod_j^r \binom{N\theta_j}{n_j}. \quad (21)$$

Условные дисперсии каждой из компонент вектора $\vec{m}$ определяются как:

$$Dm_k(\vec{n}) = \sum_{m_1=0}^{n_1} \sum_{m_2=0}^{n_2} \dots \sum_{m_r=0}^{n_r} m_k^2 \, \hbar yr(\vec{m}, \vec{n} \mid \vec{\nu}, \vec{\theta}; n) - \overline{m}_k^{\,2} =$$

$$= n\nu_k(1-\nu_k)\frac{N-n}{N-1}\binom{N}{n}^{-1}\prod_j^r \binom{N\theta_j}{n_j}, \qquad (22)$$

а элементы условной ковариационной матрицы запишутся в виде:

$$S_{kj}(\vec{n}) \equiv Cov(m_k, m_j \mid \vec{n}) =$$

$$= \sum_{m_1=0}^{n_1} \sum_{m_2=0}^{n_2} \dots \sum_{m_r=0}^{n_r} m_k \, m_j \, \hbar yr(\vec{m}, \vec{n} \mid \vec{\nu}, \vec{\theta}; n) - \overline{m}_k \, \overline{m}_j =$$

$$= -n\nu_k\nu_j \frac{N-n}{N-1}\binom{N}{n}^{-1}\prod_j^r \binom{N\theta_j}{n_j}; k, j = \overline{1,r}, k \neq j. \qquad (23)$$

Возникает вопрос: откуда нам известны значения $M_1, M_2, \dots, M_r$ (или частоты $\nu_1, \nu_2, \dots, \nu_r$)? Они нам неизвестны. И реально задача социологического исследования имеет инверсный характер: зная априори N и вектор $\theta_1, \theta_2, \dots, \theta_r$, мы из опроса получаем значения векторов $n_1, n_2, \dots, n_r$ и $m_1, m_2, \dots, m_r$. И по ним оцениваем вектор $M_1, M_2, \dots, M_r$ (и значения частот $\nu_1, \nu_2, \dots, \nu_r$), а также вычисляем погрешности этих оценок.

Распределение «квотного» опроса

Пусть квотная выборка построена по s переменным, имеющим номинальный уровень измерения, причем j-я из них имеет r_j градаций. В этом случае генеральная совокупность разбивается на $r = \prod_{j=1}^{s} r_j$ непересекающихся подмножеств численностью $N_j (k = \overline{1,r})$. Соответственно выборочный ансамбль разбивается на r стохастически независимых подвыборок численностью n_j. Введем обозначение $n_j = n_{j_1 j_2 \dots j_s}$ $(j = \overline{1,r};\ j_1 \in \overline{1,r_1}, \dots, j_s \in \overline{1,r_s})$. Тогда частоту встречаемости лиц из j-й «квотной» группы для гене-

ральной совокупности следует вычислять как величину вида $\theta_j = \prod_{j_k=1}^{s} \theta_{j_k}$.

Например, пусть квотная выборка строится по трем переменным: «пол» (мужской, женский), «уровень образования» (неполное среднее, среднее, высшее), «возраст» (молодежь, лица среднего возраста (35-55 лет), пожилые (старше 55 лет)). Тогда в зависимости от заданных значений этих переменных величина $r = 2 \times 3 \times 3 = 18$.

Сохраним смысл всех обозначений из предыдущего пункта. Тогда, при квотном отборе, случайный выбор объема n разбивается (соответственно числу квот) на r стохастически независимых отборов из выборок объема $n\theta_j$ ($j = \overline{1,r}$). Откуда следует, что вероятность получить вектор наблюдений $\vec{m}$ из лиц, соответствующих выделенным квотам, в которых зафиксирован изучаемый признак, равна:

$$\pi(\vec{m} \mid n) = \prod_{j=1}^{r} hy(m_j \mid N\nu_j, N\theta_j; n\theta_j); \vec{m} \in \Re_r^{\oplus}. \qquad (24)$$

Назовем (24) *распределением квотного отбора*. Ясно, что *это распределение не совпадает с вероятностями обнаружения фиксированного вектора* $\vec{m}$ *при случайном отборе* n наблюдений из генеральной структурированной совокупности:

$$\Pr(\vec{m} \mid n) = \sum_{n_1=m_1}^{n} \sum_{n_2=m_2}^{n-n_1} \ldots \sum_{n_r=m_r}^{n-\sum_{j=1}^{r-1} n_j} \hbar yr(\vec{m}, \vec{n} \mid \vec{\nu}, \vec{\theta}; N, n) =$$

$$= hyr(\vec{m} \mid \vec{\nu}, \vec{\theta}, N; n) \neq \pi(\vec{m} \mid n). \qquad (25)$$

Естественно, и вероятность совокупного обнаружения $m = \sum_{j=1}^{r} m_j$ наблюдений, обладающих изучаемым признаком, которая при квотном отборе равна

$$\mathrm{Pb}(m\,|\,n)=\sum_{m_1=0}^{m}\sum_{m_2=0}^{m-m_1}\ldots\sum_{m_r=0}^{m-\sum_{j=1}^{r-1}m_j}\pi(\vec{m}\,|\,n)=$$

$$=\sum_{m_1=0}^{m}hy(m_1\,|\,N\nu_1,\,N\theta_1;\,n\theta_1)\sum_{m_2=0}^{m-m_1}hy(m_2\,|\,N\nu_2,\,N\theta_2;\,n\theta_2)\ldots$$

$$\ldots\sum_{m_{r-1}=0}^{m-\sum_{j=1}^{r-2}m_j}hy(m_{r-1}\,|\,N\nu_{r-1},\,N\theta_{r-1};\,n\theta_{r-1})\;Hy(m-\sum_{j=1}^{r-1}m_j\,|\,N\nu_r,\,N\theta_r;\,n\theta_r),\quad(26)$$

не совпадает с аналогичной вероятностью при *полностью случайном* отборе:

$$\Pr(m\,|\,n)=\sum_{m_1=0}^{m}\sum_{m_2=0}^{m-m_1}\ldots\sum_{m_r=0}^{m-\sum_{j=1}^{r-1}m_j}hyr(\vec{m}\,|\,\vec{\nu},\vec{\theta},\,N;\,n)=$$

$$=hy(m\,|\,N\nu,N;\,n)\neq\mathrm{Pb}(m|n);\;\;\nu\equiv\sum_{j=1}^{r}\nu_j.\qquad(27)$$

Следует ли из соотношения (27) *тот факт, что квотный опрос некорректен для оценки частоты встречаемости заданного признака в исследуемой генеральной совокупности населения? Нет, не следует.* Математическое ожидание стохастической переменной m, подчиненной распределению квотного отбора (26), равно:

$$\overline{m}=M\left(\sum_{k=1}^{r}m_k\right)=\sum_{k=1}^{r}\overline{m}_k=\sum_{k=1}^{r}\frac{n\theta_k}{N\theta_k}\frac{N\nu_k}{}=n\sum_{k=1}^{r}\nu_k=n\nu,\qquad(28)$$

а ее дисперсия равна

$$Dm=D\left(\sum_{k=1}^{r}m_k\right)\cong\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{k=1}^{r}\theta_k\eta_k(1-\eta_k),\qquad(29)$$

где $\eta_j=M_j/N_j=\nu_j/\theta_j$, $j=\overline{1,r}$; $\sum_j\theta_j\eta_j=1.\qquad(30)$

Откуда следует, что квотная выборочная частота появления изучаемого признака является состоятельной оценкой частоты встречаемости изучаемого дихотомического признака:

$$\overline{\nu} \equiv \frac{m}{n} \underset{n \to N-0}{\to} \frac{N\nu}{N} = \nu.$$

Более того, несложно показать, что эта оценка является несмещенной и асимптотически (по n) нормальной оценкой истинной частоты ν встречаемости данного признака.

Введем величину $\overline{\nu}_j = m_j / n$; $j = \overline{1,r}$. Ее дисперсия, как следует из (24), равна

$$D\overline{\nu}_j \cong \frac{\overline{\nu}_j}{n}\left(1 - \frac{\overline{\nu}_j}{\theta_j}\right). \tag{31}$$

С учетом независимости значений m_j ($j = \overline{1,r}$), дисперсию квотной оценки частоты

$$\overline{\nu} = \sum_j^r \overline{\nu}_j \tag{32}$$

можно приближенно представить в виде:

$$D\overline{\nu} \cong \frac{1}{n}\sum_j^r \overline{\nu}_j \,(1 - \frac{\overline{\nu}_j}{\theta_j}). \tag{33}$$

Заметив, что максимум диспепсии (33) достигается при условиях $\overline{\nu}_j = \theta_j / 2$; ($j = \overline{1,r}$), по правилу «трех сигм» запишем:

$$\delta\overline{\nu} \le \frac{3}{2\sqrt{n}}\sqrt{\sum_j^r \theta_j} = \frac{3}{2\sqrt{n}}. \tag{34}$$

Сравнив (34) с (9), сделаем вывод о том, что гарантированная погрешность квотного оценивания частоты встречаемости изучаемого признака приблизительно такая же, как и при прямом случайном опросе *из неструктурированной* генеральной совокупности.

Однако есть и «минус». Варьируя выражение для $D\overline{\nu}$, легко можно увидеть, что эта величина *весьма чувствительна* даже *к*

небольшим погрешностям при построении квот $n_{j_1\, j_2 \ldots\, j_s}$. Этот факт – проявление хорошо известной в современной статистике проблемы «малой выборки», которая послужила причиной создания ряда направлений современной прикладной статистики [6] – непараметрических и робастных процедур оценивания и методов «анализа данных» в рамках концепции Дж.У. Тьюки.

Статистическое оценивание частот по случайным выборкам

Можно получить высоко точные статистические оценки частот непосредственно по случайным выборкам. Ниже изложен наиболее простой путь построения таких процедур статистического оценивания, полное обоснование которого дано в статье [30]. В работе [24] приведен более тонкий метод решения этой задачи, сохраняющий исходную идею подхода. Эта идея – *исчисления условных вероятностей для статистик бинарного отношения на множествах* [22]. Причем, по сути, она не имеет ничего общего с перевзвешиванием наблюдений, как это практикуется при «ремонте» квотных выборок.

Пусть изучается генеральная совокупность населения, мощность которой равна N, и в инструментарий исследования включены два типа вопросов, которые условно обозначим как: «содержательные», общее число вариантов ответов на которые равно p; «описательные» (обозначим их число через s), определяющие априорные классификации населения, ибо данные по эти переменным имеются в Госкомстате. Для удобства изложения будем исходить из следующих обозначений индексов:

k – номер варианта ответа на содержательный вопрос, он определяет номер соответствующей дихотомической переменной, номер признака, статистическую оценку которого мы хотим получить;

i – номер описательного вопроса, определяет i-ю априорную классификацию населения;

j – номер категории населения в i-й классификации.

При этом $k = \overline{1, p};\ i = \overline{1, s},\ j = \overline{1, r_i}$.

Объем населения, относящегося к j-й категории i-й классификации, обозначим N_{ij}. Для всех используемых априорных классификаций справедливо соотношение вида: $\forall i \in \overline{1, s} : N = \sum_{j}^{r_i} N_{ij}$. Подмножество лиц, обладающих k-м признаком, одновременно относясь к j-й категории i-й классификации, обозначим N_{ij}^{k}. Общее число жителей, обладающих k-м признаком, равно $N^{k} = \sum_{j}^{r_i} N_{ij}^{k}$.

Допустим, что опрошено n ($n << N$) респондентов. Пусть в выборку попало n_{ij} лиц, относящихся к j-й категории i-й классификации, причем k-м изучаемым признаком обладают n_{ij}^{k} из них. Общее число респондентов, имеющих k-й признак, равно: $\forall i \in \overline{1, s} : n^{k} = \sum_{j}^{r_i} n_{ij}^{k}$. Введем априорные частоты вида $\theta_{ij} \equiv N_{ij} / N$, а также частоты встречаемости k-го признака среди представителей j-й категории i-й классификации: $\nu_{ij}^{k} \equiv N_{ij}^{k} / N_{ij}$. Частота встречаемости k-го признака по населению в целом определяется в виде: $\nu^{k} \equiv N^{k} / N$. При соответствующих преобразованиях эта частота выражается в виде:

$$\forall i \in \overline{1, s} : \nu^{k} = \frac{1}{N} \sum_{j}^{r_i} N_{ij}^{k} = \sum_{j}^{r_i} \theta_{ij} \nu_{ij}^{k}. \qquad (35)$$

«Грубая» оценка частоты встречаемости k-го булевого признака среди лиц j-й категории i-й классификации имеет вид:

$$\tilde{\nu}_{ij}^{k} = n_{ij}^{k} / n_{ij}. \qquad (36)$$

Несложно показать, что оценка (36) является состоятельной, несмещенной и асимптотически (по n) нормальной оценкой истинной частоты ν_{ij}^{k}. Но, как правило, значения n_{ij}^{k} оказываются малы, что обуславливает большие погрешности оценок. Поэтому эти

оценки используются только как вспомогательные для определения частот встречаемости. Определим оценку вида:

$$\hat{\nu}_{(i)}^{k} = \sum_{j=1}^{r_i} \theta_{ij} \tilde{\nu}_{ij}^{k}. \tag{37}$$

С учетом того, что

$$Dn_{ij}^{k} \cong n_{ij}^{k}(n_{ij} - n_{ij}^{k})/n_{ij} \tag{38}$$

и ковариации величин n_{ij}^{k} и n_{il}^{k} $(l \neq j)$ вычисляется в виде

$$C_{jl}^{k(i)} \equiv Cov(n_{ij}^{k}, n_{il}^{k}) \cong -n_{ij}^{k}\, n_{il}^{k}\left(\frac{1}{n_{ij}} + \frac{1}{n_{il}}\right); l \neq j, \tag{39}$$

дисперсия оценки (37) определяется в виде:

$$D\hat{\nu}_{(i)}^{k} = \sum_{j=1}^{r_i}\left[\left(\frac{\theta_{ij}}{n_{ij}}\right)^2 Dn_{ij}^{k} + 2\sum_{l<j}^{r_i-1}\frac{\theta_{ij}\theta_{il}}{n_{ij}n_{il}} C_{jl}^{k(i)}\right] =$$

$$= \sum_{j=1}^{r_i}\theta_{ij} n_{ij}^{-1}\left[\theta_{ij}\tilde{\nu}_{ij}^{k}\left(1-\tilde{\nu}_{ij}^{k}\right) - 2\sum_{l<j}^{r_i-1}\theta_{il}\tilde{\nu}_{il}^{k}\left(1+\frac{n_{ij}}{n_{il}}\right)\right]. \tag{40}$$

Легко доказать состоятельность и несмещенность оценок $\hat{\nu}_{(i)}^{k}$ $(k = \overline{1, p};\ i = \overline{1, s})$.

Для оценки частот встречаемости «содержательных переменных» (признаков) мы использовали только одну из «описательных переменных», число которых $s > 1$. Но каждую из s оценок вида (37) можно рассматривать как некоторое *измерение* искомой частоты встречаемости k-го признака, точность которого определена выражением вида (40). Эта мысль созвучна идее Е.В. Масленникова и Ю.Н. Толстовой [13] о том, что в широком смысле любое эмпирическое исследование правомерно трактовать как некоторое *измерение* изучаемой социальной системы. Каждую оценку вида (37) можно понимать как «измерение» величины (35) в рамках i-й априорной классификации.

Изначально идея получения итоговой оценки для частоты встречаемости изучаемого признака, характеризующего интересующее

нас свойство социума, была заимствована автором из теории обработки результатов экспериментов в физике [31]. В том случае, когда некоторую величину независимо измеряют несколькими приборами (с различной точностью), итоговое значение величины вычисляется как линейная суперпозиция полученных результатов с «весами», которые определяются погрешностями измерений.

Будем рассматривать полученные «частные» оценки частоты $\hat{\nu}^k_{(i)}$ как *неравноточные* и *стохастически независимые* (что правомерно с содержательной точки зрения) *измерения* истинного значения частоты ν^k. Это позволяет, как принято в математической статистике [29], итоговую оценку частоты ν^k представить в виде линейной комбинации оценок $\hat{\nu}^k_{(i)}$:

$$\hat{\nu}^k = \sum_i^s \alpha_i \hat{\nu}^k_{(i)}. \tag{41}$$

В силу требования несмещенности итоговой оценки необходимо условие ограничения на вектор $\vec{\alpha}$ вида $\sum_i^s \alpha_i = 1$. С учетом этого требования значения компонент вектора $\vec{\alpha}$ определим из условия

$$D\hat{\nu}^k = \min(\vec{\alpha}). \tag{42}$$

Как несложно показать, решение поставленной задачи определяется в виде:

$$\alpha_i = \frac{1/D\hat{\nu}^k_{(i)}}{\sum_{j=1}^s \left(D\hat{\nu}^k_{(j)}\right)^{-1}}; (i = \overline{1,s}). \tag{43}$$

Тогда искомая итоговая оценка частоты встречаемости *k*-го признака равна:

$$\hat{\nu}^k = \left(\sum_j^s \frac{\hat{\nu}^k_{(j)}}{D\hat{\nu}^k_{(j)}}\right) \Bigg/ \left(\sum_j^s \left(D\hat{\nu}^k_{(j)}\right)^{-1}\right), \tag{44}$$

а ее дисперсия вычисляется в виде $D\hat{\nu}^k = \left(\sum_i^s \left(D\hat{\nu}^k_{(i)}\right)^{-1}\right)^{-1}$. (45)

Заметим, что все эти соотношения применимы и к результатам *квотного опроса*, поскольку он – частный случай изложенного при значениях $n_{ij} = n\theta_{ij}$.

Из формулы (45) следует, что дисперсия итоговой оценки частоты встречаемости меньше, чем минимальная из дисперсий частных оценок этой частоты вида $\hat{\nu}^k_{(i)}$; $k = \overline{1, p}$, $i = \overline{1, s}$. На практике дисперсия (45) обычно оказывается кратно (а иногда и порядково) меньше минимального из значений дисперсий $D\hat{\nu}^k_{(i)}$. Использование изложенного метода, который был положен в основу информационной технологии [32], на протяжении более 15 лет в реальных исследованиях социального, политологического и маркетингового характера показало, что при объемах выборки порядка 1 500 – 2 000 наблюдений погрешности оценок ν^k обычно составляют порядка 0,005–0,015.

ЛИТЕРАТУРА

1. *Тьюки Дж.У.* Анализ данных, вычисления на ЭВМ и математика / Пер. с англ. // Современные проблемы математики. М.: Знание, 1976.

2. *Мизес Р.* Вероятность и статистика. М.; Л.: Госиздат, 1930.

3. *Колмогоров А.Н.* Основные понятия теории вероятностей. М.: Наука, 1974.

4. *Колмогоров А.Н.* Общая теория меры и исчисление вероятностей // Труды Коммунистической академии: Разд. математ. 1929. Т. 1. С. 8–21.

5. *Орлов А.И.* Статистические методы в российской социологии (тридцать лет спустя) // Социология: методология, методы, математические модели. 2005. № 20. С. 32–53.

6. *Черепанов Е.В.* О надежности статистического анализа данных в социально-экономических исследованиях // Информатика, социология, экономика, менеджмент: Межвузовский сборник научных трудов / Под ред. С.А. Клейменова, Э.Н. Фетисова. М.: Изд-во Академии менеджмента инноваций, 2005. С. 196–213. Вып. 2. Ч. 2.

7. *Кокрен У.* Методы выборочных исследований / Пер. с англ. М.: Статистика, 1976.

8. *Бернулли Я.* О законе больших чисел / Пер. с лат.; Юбил. изд. с предисл. А.А. Маркова, А.Н. Колмогорова; Под общ. ред. Ю.В. Прохорова. М.: Наука, 1986.

9. *Крамер Г.* Математические методы статистики. М.: Мир, 1975.

10. *Косолапов М.С.* Принципы построения многоступенчатой вероятностной выборки для субъектов Российской Федерации // Социологические исследования. 1997. № 10.

11. *Толстова Ю.Н.* Анализ социологических данных: методология, дескриптивная статистика, изучение связей номинальных признаков. М.: Научный мир, 2000.

12. *Толстова Ю.Н.* Теория измерений в социологии. М.: Изд-во МГУ, 2003.

13. *Масленников Е.В., Толстова Ю.Н.* Качественная и количественная стратегии: эмпирическое исследование как измерение в широком смысле // Социология и математика: Сборник избранных трудов Ю.Н. Толстовой. М.: Научный мир, 2003. С. 198–212.

14. *Пфанцагль И.* Теория измерений / Пер. с нем. М.: Мир, 1976.

15. *Татарова Г.Г.* Основы типологического анализа в социологических исследованиях. М.: Издательский дом «Высшее образование и наука», 2007.

16. *Биркгоф Г., Барти Т.* Современная прикладная алгебра / Пер. с англ. М: Мир, 1976.

17. *Лбов Г.С.* Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981.

18. *Миркин Б.Г.* Анализ качественных признаков и структур. М.: Статистика, 1980.

19. *Дэйвид Г.* Порядковые статистики / Пер. с англ. М.: Наука,1979.

20. *Гаек Я., Шидак З.* Теория ранговых критериев / Пер. с англ. М.: Наука, 1971.

21. *Кендэл М.* Ранговые корреляции / Пер. с англ. М.: Статистика, 1975.

22. *Черепанов Е.В. и др.* Статистики бинарного отношения на множествах // Проблемы перспективного планирования и управления: Сборник научных трудов. М.: Изд-во Госплана СССР, 1990. С. 88–98.

23. *Курно О.* Основы теории шансов и вероятностей / Пер. с фр. М.: Наука, 1970.

24. *Черепанов Е.В.* Вероятностно-статистические основы прикладной социологии и маркетинговых исследований. М.: Изд-во Академии менеджмента инноваций, 2006.

25. *Кендалл М., Стюарт А.* Теория распределений / Пер. с англ. М.: Наука, 1966.

26. *Миттаг Х.-Й., Ринне Х.* Статистические методы обеспечения качества / Пер. с нем. М.: Машиностроение, 1995.

27. *Янке Е. и др.* Специальные функции / Пер. с англ. М.: Наука, 1977.

28. *Люк Ю.* Специальные математические функции и их аппроксимации. М.: Мир, 1980.

29. *Кендалл М., Стюарт А.* Статистические выводы и связи / Пер. с англ. М.: Наука, 1976.

30. *Азаров С.*В., *Черепанов Е.В.* Регрессионные методы статистического оценивания в социальных исследованиях // Математические методы и компьютерные технологии в маркетинговых и социальных исследованиях: Сборник научных работ / Под общ. ред. С.А. Клейменова. М.: Изд-во Академии менеджмента инноваций, 2004. С. 56–72.

31. *Свешников А.А.* Основы теории ошибок. Л.: Изд-во ЛГУ, 1972.

32. *Азаров С.В., Пашин Ю.А., Черепанов Е.В.* Современные компьютерные технологии в социальных исследованиях // Безопасность Евразии. 2005. № 1. С. 264–281.