Е. В. Черепанов

Стохастические методы анализа данных выборочных маркетинговых и социальных обследований[1]

Любые количественные выводы в маркетинге потребительских рынков и прикладной социологии основаны на асимптотических свойствах выборочных частот. Для преодоления проблемы неоднородности населения во всем мире используют метод «квотных выборок», отражающих по основным категориям структуру генеральной совокупности. В работе предложен метод статистического анализа данных о конечных структурированных множествах, которые получены на основе случайного отбора. Метод основан на исчислении условных вероятностей для статистик бинарных отношений на множествах «наблюдения — дихотомические признаки». По сравнению с квотными методами, предложенный подход значительно повышает точность оценок по населению (покупателям, избирателям) в целом и позволяет получить оценки частот по категориям населения для любых априорных классификаций.

Ключевые слова: структурированное конечное множество, выборочный метод, дихотомические (булевы) признаки, статистические оценки, случайная выборка, квотная выборка, гипергеометрическое распределение, маркетинг потребительских рынков, прикладные социологические исследования.

JEL classification: C13, C81, C83.

1. Введение

Выборочный метод заложен в основу любых методик маркетинговых и прикладных социологических исследований. Заметим, что здесь и далее термин «маркетинг» используется в узком смысле, как изучение *массового* потребительского рынка. С математической точки зрения выборочные обследования (при отборе из *однородной* совокупности) подчинены гипергеометрическому распределению (ГГР) вероятностей, что, насколько это известно автору, в русскоязычной литературе было впервые отмечено в монографии (Кокрен, 1976).

[1] *Примечание от редакции*. Статья затрагивает очень важную для приложений тему теоретико-вероятностного обоснования формирования случайных неоднородных выборок. Поэтому редколлегия журнала решила опубликовать эту статью, несмотря на отдельные критические замечания рецензентов. О двух таких замечаниях мы решили проинформировать читателя. Во-первых, автор подвергает чрезмерной критике широко распространенный и оправдавший себя в социологических исследованиях метод, основанный на квотных выборках. Второе замечание относится к точности полученных автором оценок дисперсий частот изучаемого признака. Так, высказанное автором соображение, что «измерения частоты встречаемости качественного признака с помощью практически не связанных между собой номинальных шкал должны слабо коррелировать», часто оказывается сомнительным в практической работе и зачастую требует дополнительной коррекции.

По сути, любые выборочные методики базируется на законе больших чисел, согласно которому (в форме теоремы Я. Бернулли (Бернулли, 1986)) выборочная частота встречаемости булевого признака в серии независимых наблюдений сходится по вероятности к его истинной частоте встречаемости. Сложность состоит в том, что, кроме случайности и независимости наблюдений, требуется априорная *однородность* наблюдений. А население — это структурированное (по многим номинальным шкалам) множество. В этой связи, при разумных объемах выборки (например, 2–3 тыс. *случайно опрошенных* респондентов), различие в структурах выборки и генеральной совокупности могут сильно испортить точность оценки частоты встречаемости изучаемого признака (Мхитарян, Черепанов, 2006; Черепанов, 2007б).

В принципе существует лишь два решения этой проблемы: 1) при расчетах математически строго учесть различия в структурах выборки и генеральной совокупности; и 2) так подобрать выборку, чтобы ее структура по основным классификациям (пол, возраст, образование, национальность и т. п.) дублировала бы генеральную совокупность (построить так называемую «*квотную*» выборку). Поскольку в 30-е гг. прошлого века, когда зародились массовые опросы населения, вычислительной техники не существовало, то у пионеров эмпирической социологии фактически выбора не было: раз считать условные вероятности не на чем, будем создавать квотные выборки.

Эксплуатируя квотные выборки, почти век развивались маркетинг рынков и прикладная социология. За это время методы математической работы с социально-экономическими и социологическими данными достигли серьезного уровня (см., например, (Толстова, 2000; Айвазян, Мхитарян, 2001; Орлов, 2006)). Но в эмпирических исследованиях продолжается применение «квотных технологий», в частности, классических методов математической статистики, которые были заимствованы из анализа однородных данных в естественнонаучных областях знания. А в части структурированности населения «молча» предполагается, что все связанные с ней проблемы разрешены на этапе проведения квотного опроса населения.

Кроме того, формирование, «ремонт» и поддержание квотных выборок — дело дорогостоящее и связанное с большими трудозатратами (Косолапов, 1997). Однако создание точных и дешевых методов работы со случайными выборками из структурированных (несколькими номинальными шкалами) совокупностей, которые используют современные компьютерные технологии (Азаров и др., 2005), описываются в терминах классической теории вероятностей. Эти методы, основанные на исчислении статистик бинарных отношений на множествах (Колочков и др., 1990), используют обобщения гипергеометрического распределения вероятностей.

2. Многомерные обобщения структурированного ГГР

Статистические процедуры, на которых базируется выборочный метод в социологических исследованиях, основаны на ГГР (Справочник по теории вероятностей…, 1978, п. 6.1.5). Пусть задана генеральная совокупность, представляющая собой население (избиратели, покупатели и т. д.), состоящая из N человек ($N \gg 1$). Среди населения существует M человек, обладающих интересующим нас дихотомическим признаком (состоят в данной партии, являются клиентами некоторого пенсионного фонда, пользуются изучаемой страховой услугой и т. п.). Производится *случайная выборка* респондентов объема n. Вероятность того,

что в выборку попадет *ровно m* лиц, обладающих изучаемым дихотомическим признаком ($0 \le m \le n$), задается формулой:

$$\Pr\{m \mid n\} = hy(m \mid M, N; n) = \binom{N}{n}^{-1} \binom{M}{m} \binom{N-M}{n-m}, \qquad (2.1)$$

где $\binom{N}{n} = \frac{N!}{(N-n)!\,n!}$, $\Pr\{\cdot\}$ обозначает вероятность события $\{\cdot\}$, а $hy(\cdot)$ — стандартное обозначение ГГР (Миттаг, Ринне, 1995).

Используя свойства гамма-функции (Янке и др., 1977, п. V. 3), из (2.1), несложно получить

$$hy(m \mid N, M; n) = \left(1 + \frac{m}{N-n+1}\right)\left(\prod_{k=1}^{M}\left(1 - \frac{n-m}{N-k+1}\right)\right)\prod_{k=1}^{m}\frac{(M-k+1)(n-k+1)}{l(N-n+k+1)}.$$

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

Отметим, что ГГР зародилась в задачах анализа качества массовой продукции (Миттаг, Ринне, 1995). Но сегодня многомерные обобщения ГГР могут быть широко использованы 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 при актуарных расчетах в страховом деле.

Пусть изучается генеральная совокупность населения, мощность которой равна *N*. Для маркетингового или социологического опроса составлен инструментарий из некоторого числа «содержательных вопросов», общее число *вариантов ответов* на которые равно *p*. При опросе используются *s* априорных классификаций, данные по которым имеются в Росстате (обычно, это данные последней переписи населения).

В дальнейшем будем обозначать:

- индексом *k* — номер варианта ответа на содержательный вопрос анкеты, иначе говоря, *k* определяет номер соответствующего *булевого признака*, характеризующего наблюдения изучаемой совокупности;
- индексом *i* — номер априорной классификации (номинальной шкалы), данные по которой есть в Госкомстате;
- индексом *j* — номер социально-демографической категории населения (покупателей, электората), определенной *i*-ой априорной классификацией.

Итак, если прямо не оговорено иное, везде далее $k = 1,...,p$; $i = 1,...,s$; $j = 1,...,r_i$.

Например, $k = 45$ — намерение купить автомобиль «Форд Фокус», $i = 4$ — классификация по возрасту, $j = 3$ — лица в возрасте 45–60 лет. В этом случае запись $N_{4,3}^{45} = 5200$ означает, что в заданном регионе существует 5200 лиц из указанного возрастного диапазона, желающих купить «Форд Фокус».

Общее число жителей, относящихся к *j*-ой категории *i*-ой классификации, обозначим N_{ij}. Для всех априорных классификаций населения справедливо соотношение $N = \sum_{j=1}^{r_i} N_{ij}$, $i = 1,...,s$.

Этот очевидный факт объясняется тем, что каждая номинальная шкала (классификация) задает разбиение (непересекающееся покрытие) населения. Мощность подмножества лиц, обладающих *k*-ым «содержательным» признаком, одновременно относясь к *j*-ой категории *i*-ой классификации, обозначим N_{ij}^k. Ясно, что общее число жителей, обладающих *k*-ым признаком, для любой априорной классификации (при любом *i*) равно $N^k = \sum_{j=1}^{r_i} N_{ij}^k$.

В ходе случайного опроса было проинтервьюировано n ($n < N$) респондентов. Пусть в выборку попало n_{ij} лиц, относящихся к *j*-ой категории *i*-ой классификации, причем *k*-ым изучаемым признаком обладают n_{ij}^k из них. Общее число респондентов, имеющих *k*-ый признак, равно $n^k = \sum_{j=1}^{r_i} n_{ij}^k$.

Введем априорные частоты вида θ_{ij}, определяющие доли численности *j*-ой категории *i*-ой классификации среди всего изучаемого населения:

$$\theta_{ij} = N_{ij} / N;\ \forall\, i = 1,\ldots,s:\ \sum_{j=1}^{r_i} \theta_{ij} = 1\ ;\ \text{причем}\ \ \sum_{j=1}^{r_i} n_{ij} = n. \qquad (2.2)$$

Введем векторные обозначения:

$$\vec{n}_i = (n_{i1}, n_{i2}, \ldots, n_{ir_i}) \in \mathbb{R}_{r_i}^+;\ \vec{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ir_i}) \in \mathbb{R}_{r_i}^+.$$

Подчеркнем, что компоненты этих векторов известны: значения n_{ij} непосредственно по структуре полученной в ходе опроса случайной выборки, а значения априорных частот θ_{ij} — данные Росстата.

Вероятность $\Pr\{\vec{n}_i \mid n\}$ того, что случайная выборка объема *n* по *i*-ой классификации имеет структуру $\vec{n}_i$, определяется *многомерным* ГГР вида

$$\Pr\{\vec{n}_i \mid n\} = hyr_i\,(\vec{n}_i \mid \vec{\theta}_i, N; n) = \binom{N}{n}^{-1} \prod_{j=1}^{r_i} \binom{N\theta_{ij}}{n_{ij}}. \qquad (2.3)$$

Например, пусть *i* — классификация населения региона по национальностям, тогда: θ_{i1} — доля русских среди населения, θ_{i2} — украинцев, …, θ_{ir_i} — армян. Вероятность того, что в случайную выборку объема *n* попадет n_{i1} русских, n_{i2} украинцев, …, n_{ir_i} армян, описывается распределением (2.3). Введем в рассмотрение частоты встречаемости *k*-го «содержательного» признака среди представителей *j*-ой категории *i*-ой классификации ν_{ij}^k: $\nu_{ij}^k = N_{ij}^k / N_{ij}$. Частота встречаемости *k*-го признака по населению в целом определяется в виде $\nu^k = N^k / N$. С помощью категорий априорной классификации населения эта частота выражается в виде

$$\forall\, i \in \overline{1,s} \qquad \nu^k = \frac{1}{N} \sum_{j=1}^{r_i} N_{ij}^k = \sum_{j} \theta_{ij} \nu_{ij}^k \le 1\ . \qquad (2.4)$$

Введем вектор вида: $\vec{n}_i^k = (n_{i1}^k, n_{i2}^k, \ldots, n_{ir_i}^k) \in \mathbb{R}_{r_i}^+$. Вероятность $\Pr\{\vec{n}_i^k \mid \vec{n}_i\}$ того события, что в случайной выборке объема *n* со структурой по *i*-ой классификации вида $\vec{n}_i$ *k*-ый признак будет зафиксирован в виде вектора $\vec{n}_i^k$, определится как:

$$\Pr\{\vec{n}_i^k \mid \vec{n}_i\} = \prod_{j=1}^{r_i} hy\,(n_{ij}^k \mid N\theta_{ij}\nu_{ij}^k, N\theta_{ij}; n_{ij}) = hnr_i\,(\vec{n}_i^k \mid \vec{\theta}_i, \vec{\nu}_i^k; N, \vec{n}_i) =$$

$$= \prod_{j=1}^{r_i} \binom{N\theta_{ij}}{n_{ij}}^{-1} \binom{N\theta_{ij}\nu_{ij}^k}{n_{ij}^k} \binom{N\theta_{ij}(1-\nu_{ij}^k)}{n_{ij}-n_{ij}^k}; \quad \vec{n}_i, \vec{n}_i^k \in N_{r_i}. \tag{2.5}$$

Распределение $hnr_i\,(\vec{n}_{ij}^k \mid \vec{\theta}_i, \vec{\nu}_i^k; N, \vec{n}_i)$, которое назовем *условным распределением структурированной выборки 1-го рода*, определяет распределение лиц с *k*-ым изучаемым признаком в выборке по категориям *i*-ой классификации при некоторой заданной структуре выборки (по этой классификации).

Далее, используя (2.3) и (2.5), вероятность того, что в случайной выборке объема *n* окажется: а) по *i*-ой классификации структура $\vec{n}_i$, и б) *k*-ый признак будет зафиксирован в виде вектора $\vec{n}_i^k$, определяется как:

$$\Pr\{\vec{n}_i^k, \vec{n}_i \mid n\} = \Pr\{\vec{n}_i^k \mid \vec{n}_i\} \Pr\{\vec{n}_i \mid n\} =$$

$$= \hbar yr_i\,(\vec{n}_i^k, \vec{n}_i \mid \vec{\nu}_i^k, \vec{\theta}_i; N; n) = \binom{N}{n}^{-1} \prod_{j=1}^{r_i} \binom{N\theta_{ij}\nu_{ij}^k}{n_{ij}^k} \binom{N\theta_{ij}(1-\nu_{ij}^k)}{n_{ij}-n_{ij}^k}. \tag{2.6}$$

Распределение вида (2.6) назовем *многомерным структурированным гипергеометрическим распределением* (МСГГР). Очевидно, что одномерное СГГР определится в виде

$$\hbar y(n_{ij}^k, n_{ij} \mid N, \nu_{ij}^k, \theta_{ij}; n) = hy(n_{ij}^k \mid N\theta_{ij}\nu_{ij}^k, N\theta_{ij}; n_{ij})\, hy(n_{ij} \mid N\theta_{ij}, N; n) =$$

$$= \binom{N}{n}^{-1} \binom{N(1-\theta_{ij})}{n-n_{ij}} \binom{N\theta_{ij}\nu_{ij}^k}{n_{ij}^k} \binom{N\theta_{ij}(1-\nu_{ij}^k)}{n_{ij}-n_{ij}^k}. \tag{2.7}$$

Теперь определим необходимое для построения статистических процедур *условное распределение структурированной выборки 2-го рода*:

$$\Pr\{\vec{n}_i \mid \vec{n}_i^k\} = \frac{\Pr\{\vec{n}_i^k, \vec{n}_i \mid n\}}{\Pr\{\vec{n}_i^k \mid n\}} = \hbar r_i\,(\vec{n}_i \mid \vec{\nu}_i^k, \vec{\theta}_i, N; \vec{n}_i^k) =$$

$$= \binom{N(1-\nu^k)}{n-n^k}^{-1} \prod_{j=1}^{r_i} \binom{N(1-\theta_{ij})}{n-n_{ij}} \binom{N\theta_{ij}(1-\nu_{ij}^k)}{n_{ij}-n_{ij}^k}. \tag{2.8}$$

Распределение $\hbar r_i\,(\vec{n}_i \mid \vec{\nu}_i^k, \vec{\theta}_i, N; \vec{n}_i^k)$ определяет вероятность конкретной структуры выборки (по данной классификации) при заданной структуре выборки по *k*-ому изучаемому признаку. Одномерный вариант этого условного распределения имеет вид:

$$\hbar(n_{ij} \mid \nu_{ij}^k, \theta_{ij}, N; n_{ij}^k) = hy[\, n_{ij} - n_{ij}^k \,|\, N(1-\theta_{ij}\nu_{ij}^k), N\theta_{ij}(1-\nu_{ij}^k); n-n_{ij}^k] =$$

$$= \binom{N(1-\theta_{ij}\nu_{ij}^k)}{n-n_{ij}^k}^{-1} \binom{N(1-\theta_{ij})}{n-n_{ij}} \binom{N\theta_{ij}(1-\nu_{ij}^k)}{n_{ij}-n_{ij}^k}. \tag{2.9}$$

3. Квотные методы выборочных обследований

Для прояснения сути вопроса рассмотрим сначала *однородные данные*, подчиненные *одномерному* ГГР (2.1). Зная вид распределения, несложно вычислить значения границ доверительных интервалов для «прямых оценок» частот встречаемости дихотомических признаков

$$\hat{\nu} = m/n \tag{3.1}$$

при заданном уровне доверительной вероятности. Но этот процесс достаточно трудоемок. Поэтому, с удовлетворительной точностью ограничимся значениями оценок погрешностей «сверху» для оценок частот (3.1). Используя выражение для дисперсии ГГР (Справочник по теории вероятностей…, 1978, п. 6.1.6), имеем

$$Dm = n\nu(1-\nu)\frac{1-n/N}{1-1/N} \cong n\nu(1-\nu). \tag{3.2}$$

Как правило, при *массовых* маркетинговых и социологических опросах $n \ll N$. Это позволяет, используя правило «трех сигм», представить (на уровне доверительной вероятности не менее 0.99) гарантированную оценку погрешности частоты встречаемости $\hat{\nu}$ в виде

$$\delta \cong 3\sqrt{D\hat{\nu}} \cong 3\sqrt{\hat{\nu}(1-\hat{\nu})/n} \le \frac{3}{2\sqrt{n}}. \tag{3.3}$$

В формуле (3.3) учтено, что максимум дисперсии $D\hat{\nu}$ достигается при значении $\hat{\nu} = 0.5$. Используя это соотношение, вычисляются значения гарантированных погрешностей для «прямых» оценок частот встречаемости дихотомического признака, подчиненного ГГР, в зависимости от n.

Интересны и «обратные» оценки: каковы должны быть объемы выборки для заданных уровней гарантированной погрешности? Из неравенства (3.3) получаем приближение:

$$n \cong 9/(4\delta^2). \tag{3.4}$$

Соответствующие данные приведены в табл. 1.

Таблица 1. Необходимые объемы выборки для заданных уровней гарантированной погрешности «прямых» оценок частот встречаемости дихотомических признаков

δ	0.005	0.01	0.02	0.03	0.04	0.05	0.10	0.15
n	90000	22500	5600	2500	1400	600	225	100

Заметим, что для (традиционных в социологии и маркетинге) объемов выборки порядка 1.5–2 тыс. респондентов гарантированная погрешность частоты примерно равна 3.5%, как обычно и указывается в публикациях. Но для точности оценок в 2% нужно уже порядка 5.5 тыс. наблюдений, а гарантия погрешности в 1% потребует опроса 22.5 тыс. респондентов.

Важно и то, что если нужно сделать статистические выводы по некоторой немногочисленной категории населения, то численность этой категории в репрезентативной выборке должна составлять (при разумном пороге точности в 5%) не менее 600 (!) человек. Это значит, например, что для категории, которая составляет 5% населения (скажем, «военнослужащие в Пермском крае» или «таджики в Хакасии») потребуется квотная выборка (подробнее см. далее) порядка 12 тыс. человек. *Практически методом квотного опроса это нереализуемо*. В данном примере потребуется объем квотной выборки $n = 12/0.05 = 240$ тыс. человек. Следовательно, для анализа структуры общественного мнения нужны отдельные исследования *для каждой такой категории населения*. *Это существенный вывод*: для оценки частот встречаемости качественных признаков *по категориям населения* не смогут помочь самые совершенные стандартные пакеты программ.

Хотя использование квотных методик в исследованиях общественного мнения и предпочтений потребителей сегодня носят тотальный характер, не было найдено ни одной публикации, в которой формально обосновывалась бы корректность применения статистических методов на квотных выборках (за исключением статьи (Черепанов, 2007в)). Но по своему построению *квотные выборки в строгом понимании не являются случайными*. Следовательно, правомерность их применения и корректность полученных на них статистических выводов требует обоснования.

Рассмотрим суть квотного отбора. Пусть, как и ранее, население имеет априорные классификации по s номинальным шкалам, причем j-ая шкала имеет r_j категорий. Тогда генеральная совокупность разбивается на $r = \prod_{j=1}^{s} r_j$ непересекающихся подмножеств («квот») численностью N_l $(l = 1,...,r)$. Частота встречаемости лиц l-ой «квотной группы» из генеральной совокупности, обозначаемая η_l $(l = 1,...,r)$, вычисляется как

$$\eta_l = \prod_{k=1}^{s} \theta_{jl_k}. \qquad (3.5)$$

Пример. Пусть построение квотной выборки производится по трем шкалам наименований: «пол», «уровень образования», «возраст». Первая шкала имеет два значения ($r_1 = 2$). Вторая шкала ($k = 2$) имеет три значения ($r_2 = 3$): «неполное среднее», «среднее» и «высшее» образование. Третья шкала ($k = 3$) имеет четыре значения ($r_3 = 4$): «молодежь» (до 30 лет), «лица среднего возраста» (31–45 лет), «зрелые люди» (46–60 лет) и «пожилые» (старше 60 лет). Тогда $r = 2 \cdot 3 \cdot 4 = 24$. Заметим, что если добавить четвертую классификацию, например, «условия проживания» с пятью категориями («мегаполис» (более 1 млн жителей), «город» (100 тыс.–1 млн жителей), «городок» (до 100 тыс. жителей), «поселок» городского типа, «сельская местность»), то число «квотных групп» возрастет до $r = 5 \cdot 24 = 120$. Формирование такой квотной выборки на практике становится крайне трудоемким занятием.

Если же добавить пятую классификацию, скажем, «национальность», например, с 15 значениями («русск.», «укр.»,…, «калмык», «проч.»), то число «квот» возрастет до $r = 15 \cdot 120 = 1800$. И формирование такого квотного выборочного ансамбля становится уже просто нереальным. В этой связи, по крайней мере в России, при квотных обследованиях обычно ограничиваются тремя-четырьмя классификациями (как правило, пол, возраст, образование, иногда — регион проживания или условия проживания).

При квотном отборе псевдослучайная выборка объема n формируется (соответственно числу квот) путем r стохастически независимых случайных отборов (по каждой из квот) объемами $n\eta_l$. Далее, пусть среди N_l лиц, входящих в l-ую квотную группу, ровно $M_l = N_l\nu_l$ лиц обладают изучаемым дихотомическим признаком. Общее число лиц генеральной совокупности, обладающих этим дихотомическим признаком, равно $M = N\sum_{l=1}^{r}\eta_l\nu_l$. Ясно, что общая «частота встречаемости» этого признака равна $\nu = M/N = \sum_{l=1}^{r}\eta_l\nu_l$. В прикладных задачах, как правило, значения η_l $(l=1,...,r)$ известны, а значения ν_l $(l=1,...,r)$ и ν неизвестны.

Несложно понять, что вероятность получить вектор наблюдений $\vec{m} = \{m_1, m_2, ..., m_r,\}$ из лиц, обладающих изучаемым булевым признаком и входящих в соответствующую «квотную группу», равна

$$\pi(\vec{m} \mid n) = \prod_{j=1}^{r} hy(m_j \mid N\eta_j\nu_j, N\eta_j; n\eta_j);\ \vec{m} \in \mathbb{R}_r^+. \tag{3.6}$$

Назовем (3.6) *структурированным распределением квотного отбора* (СРКО). Отсюда ясно, что вероятность совокупного обнаружения $m = \sum_{j=1}^{r} m_j$ наблюдений, обладающих изучаемым признаком, при квотном отборе определяется выражением, которое назовем *распределением квотного отбора* (РКО)

$$\begin{aligned}\pi(m|n) = &\left[\prod_{j=1}^{r}\binom{N\eta_j}{n\eta_j}^{-1}\right]\sum_{m_r=0}^{m}\binom{N\eta_r\nu_r}{m_r}\binom{N\eta_r(1-\nu_r)}{n\eta_r-m_r}\cdot\ldots\cdot\sum_{m_{r-1}=0}^{m-m_r}\binom{N\eta_{r-1}\nu_{r-1}}{m_{r-1}}\binom{N\eta_{r-1}(1-\nu_{r-1})}{n\eta_{r-1}-m_{r-1}}\cdot\ldots\\ &\ldots\cdot\sum_{m_2=0}^{m-\sum_{i=3}^{r}m_i}\binom{N\eta_2\nu_2}{m_2}\binom{N\eta_2(1-\nu_2)}{n\eta_2-m_2}\binom{N\eta_1\nu_1}{m-\sum_{i=3}^{r}m_i}\binom{N\eta_1(1-\nu_1)}{n\eta_1-m+\sum_{i=3}^{r}m_i}.\end{aligned} \tag{3.7}$$

По-видимому, путем комбинаторных преобразований РКО можно придать вид, более обозримый, чем (3.7). Но в силу произвольности значений частот ν_j $(j=1,...,r)$ очевидно, что *нельзя* привести (3.7) к ГГР, определяющему случайный отбор

$$hy(m \mid N\nu, N; n) = \binom{N}{n}^{-1}\binom{N\nu}{m}\binom{N(1-\nu)}{n-m}. \tag{3.8}$$

Следует ли отсюда, что *квотный опрос со стохастической точки зрения некорректен для оценки частоты встречаемости заданного признака в исследуемой генеральной совокупности? Нет, не следует.* В статье (Черепанов, 2007в) показано, что математическое ожидание случайной переменной $m = \sum_{k=1}^{r} m_k$, подчиненной РКО, равно $n\nu$, а ее дисперсия

асимптотически (по n) стремится к нулю. Следовательно, квотная выборочная частота появления изучаемого дихотомического признака является несмещенной и состоятельной оценкой истинной частоты встречаемости этого дихотомического признака.

4. Погрешности квотного метода в социальных работах

Введем величину

$$\overline{\nu}_l = m_l / n\,, \quad l = 1,...,r\,. \tag{4.1}$$

Ее дисперсия, учитывая, что отбор по каждой квоте подчинен соответствующему гипергеометрическому распределению, приближенно равна

$$D\overline{\nu}_l \cong \frac{\overline{\nu}_l}{n}\left(1 - \frac{\overline{\nu}_l}{\eta_l}\right). \tag{4.2}$$

С учетом очевидной стохастической независимости значений m_l $(l = 1,...,r)$, дисперсию квотной оценки «суммарной» частоты вида

$$\overline{\nu} = \sum_{l=1}^{r} \overline{\nu}_l \tag{4.3}$$

можно приближенно представить как

$$D\overline{\nu} \cong \frac{1}{n}\sum_{l=1}^{r} \overline{\nu}_l\,(1 - \overline{\nu}_l/\eta_l). \tag{4.4}$$

Заметив, что максимум дисперсии (4.4) достигается при условиях $\overline{\nu}_l = \eta_l / 2 \quad (l = 1,...,r)$, по правилу «трех сигм» запишем:

$$\delta\overline{\nu} \le \frac{3}{2\sqrt{n}}\sqrt{\sum_{l=1}^{r}\eta_l} = \frac{3}{2\sqrt{n}}\,. \tag{4.5}$$

Сравнивая (4.5) с (3.3), видим, что гарантированная погрешность квотного оценивания частоты встречаемости дихотомического признака имеет тот же порядок, что и погрешности оценивания частот при прямом случайном опросе *из неструктурированной* генеральной совокупности.

5. Статистические оценки частот встречаемости булевых признаков

На основе обобщений ГГР, описанных в п. 2, возможны различные виды состоятельных оценок частот встречаемости дихотомических признаков как по населению в целом, так и по его социально-демографическим категориям (Черепанов, 2006, 2007а, 2008). Ниже приведен простой метод, позволяющий получить достаточно точные оценки частот.

Используя условное распределение случайной выборки 1-го рода (2.5) вида

$$hnr_i\left(\vec{n}_i^k \mid \vec{\theta}_i, \vec{\nu}_i^k; N, \vec{n}_i\right) = \prod_{j=1}^{r_i} hy\left(n_{ij}^k \mid N\theta_{ij}\nu_{ij}^k, N\theta_{ij}; n_{ij}\right),$$

«грубую» оценку частоты встречаемости k-го булевого признака среди лиц j-ой категории i-ой классификации запишем в виде

$$\tilde{\nu}_{ij}^k = n_{ij}^k / n_{ij}. \tag{5.1}$$

Несложно показать, что (5.1) является состоятельной и несмещенной оценкой частоты ν_{ij}^k. Но на практике значения n_{ij} и n_{ij}^k оказываются, зачастую, малы, что обуславливает большие погрешности оценок (4.1). Поэтому эти оценки используются только как *вспомогательные*. Определим оценку вида

$$\hat{\nu}_{(i)}^k = \sum_{j=1}^{r_i} \theta_{ij}\tilde{\nu}_{ij}^k, \tag{5.2}$$

рассмотрев условное распределение (2.5) структурированной выборки 1-го рода $hnr_i\left(\vec{n}_i^k \mid \vec{\theta}_i, \vec{\nu}_i^k; N, \vec{n}_i\right)$. Дисперсия случайной величины n_{ij}^k приближенно равна

$$Dn_{ij}^k \cong n_{ij}\tilde{\nu}_{ij}^k(1-\tilde{\nu}_{ij}^k)(1-n_{ij}/N_{ij}), \quad N \gg 1. \tag{5.3}$$

Следовательно, дисперсия оценки $\tilde{\nu}_{ij}^k$ выражается в виде

$$D\tilde{\nu}_{ij}^k \cong \tilde{\nu}_{ij}^k(1-\tilde{\nu}_{ij}^k)\,(n_{ij}^{-1} - N_{ij}^{-1}). \tag{5.4}$$

Поскольку ковариации случайных величин n_{ij}^k и n_{il}^k $(l \neq j)$ для распределения структурированной выборки 1-го рода равны нулю, то дисперсия оценки (5.2) выглядит как

$$D\hat{\nu}_{(i)}^k = \sum_{j=1}^{r_i} \theta_{ij}^2 D\tilde{\nu}_{ij}^k. \tag{5.5}$$

Тривиально показать состоятельность и несмещенность оценок $\hat{\nu}_{(i)}^k$.

Каждую из s оценок вида (5.2) можно рассматривать как некоторое *неравноточное измерение* искомой частоты встречаемости k-го признака, погрешность которого определена дисперсией вида (5.5). Уместно отметить, что идея получения итоговой оценки частоты встречаемости изучаемого признака в виде линейной суперпозиции ее неравноточных измерений (Свешников, 1972) соответствует естественнонаучной традиции обработки результатов экспериментов (Мудров, Кушко, 1976).

Будем рассматривать «частные» оценки частоты $\hat{\nu}_{(i)}^k$ как *неравноточные измерения* истинного значения частоты ν^k. Итоговую оценку частоты ν^k представим в виде

$$\hat{\nu}^k = \sum_{i=1}^{s} \alpha_i \hat{\nu}_{(i)}^k. \tag{5.6}$$

Ее дисперсия имеет вид $D\hat{\nu}^k = \sum_{i=1}^{s} \alpha_i^2 D\hat{\nu}_{(i)}^k + \sum_{i=1}^{s}\sum_{j=1}^{s} \alpha_i\alpha_j C_{ij}^k$, где $C_{ij}^k = \mathrm{Cov}(\hat{\nu}_{(i)}^k, \hat{\nu}_{(j)}^k)$. Но в статье (Азаров, Черепанов, 2004), основываясь на вычислениях ковариаций по методу из рабо-

ты (Вучков и др., 1987), показано, что значения $\left|C_{ij}^{k}\right|$, как правило, на порядок меньше, чем значения $D\hat{\nu}_{(i)}^{k}$. Содержательно это ясно: измерения частоты встречаемости качественного признака с помощью практически не связанных между собой номинальных шкал должны слабо коррелировать. В этой связи величины $\hat{\nu}_{(i)}^{k}$ в первом приближении можно считать статически независимыми.

Для несмещенности оценки (5.6), необходимо ограничение на вектор $\vec{\alpha}$ вида $\sum_{i=1}^{s}\alpha_i = 1$.

С учетом этого требования, значения компонент вектора $\vec{\alpha}$ можно определить из критерия

$$D\hat{\nu}^{k} \to \min_{\vec{\alpha}}. \tag{5.7}$$

Решение этой задачи, в предположении стохастической независимости «неравноточных измерений» $\hat{\nu}_{(i)}^{k}$, находится в виде оценки

$$\hat{\nu}^{k} = \left(\sum_{j=1}^{s}\frac{\hat{\nu}_{(j)}^{k}}{D\hat{\nu}_{(j)}^{k}}\right) \Bigg/ \left(\sum_{j=1}^{s}(D\hat{\nu}_{(j)}^{k})^{-1}\right), \tag{5.8}$$

дисперсия которой равна

$$D\hat{\nu}^{k} = \left(\sum_{i=1}^{s}\left(D\hat{\nu}_{(i)}^{k}\right)^{-1}\right)^{-1}. \tag{5.9}$$

Являясь средним гармоническим дисперсий вспомогательных оценок, дисперсия оценки (5.9) заведомо меньше минимального значения этих дисперсий. Заметим, что *все соотношения этого пункта применимы и к результатам квотного опроса*, поскольку он представляет собой частный случай изложенного при значениях $n_{ij} = n\theta_{ij}$.

Пример: прогноз итогов голосований на Съезде народных депутатов

Приведем пример из практики автора. В 1992 году администрацией Президента РФ было решено пригласить на очередной VII Съезд народных депутатов России, как это практикуется в Конгрессе США, семь коллективов социологов, шесть из которых являлись наиболее известными социологическими центрами РФ. Седьмой приглашенной организацией был Институт системных исследований и социологии (ИСИС) — частная структура, директором которой тогда был автор.

На Съезде остро встал вопрос, который был крайне актуален для администрации Президента РФ: имеет ли шансы Е. Т. Гайдар, еще возглавлявший правительство России, сохранить свой пост. Кураторы социологических работ на Съезде А. Н. Лифшиц (впоследствии ставший министром финансов РФ) и И. Г. Яковлев (ныне профессор Московского городского университета управления) задали этот вопрос социологам. Шесть команд, занимавшихся описанием позиций депутатов, не смогли дать вразумительный ответ о шансах Е. Т. Гайдара.

ИСИС через час после поступления вопроса дал ответ: «за» сохранение поста Е. Т. Гайдаром будут *470 депутатов плюс-минус 6 голосов*. Это означало, что действующий премьер ни в коем случае не сможет получить поддержку большинства депутатов (которая составляла 521 голос). Через сутки процедура тайного голосования дала результат: *за сохранение поста премьер-министра Е. Т. Гайдаром было отдано 467 голосов народных депутатов РФ*.

Подход состоял в следующем. Все команды социологов получили распечатки поименных голосований депутатов на предыдущих съездах. Нашими коллегами эти распечатки использовались для сопоставительного анализа позиций депутатского корпуса. А команда ИСИС отобрала 125 голосований по *важнейшим* вопросам и использовала их в качестве номинальных шкал (априорных классификаций) со значениями: «за», «отсутствовал» и «прочее» (позиции «против» и «воздержался» были равнозначны с точки зрения итогов голосования). В результате каждый депутат обрел «опросный паспорт» из 125 номинальных шкал, который использовался при решении задачи прогнозирования итогов голосований.

Первым вопросом, который задавался каждому из опрашиваемых депутатов (для его идентификации в базе данных), был: «Пожалуйста, представьтесь». Ответив, респондент автоматически «заполнял» «социологический паспорт», априорные частоты которого были известны из распечатки результатов предыдущих голосований. Таким образом, опросив всего лишь около 40 депутатов (каждый из которых имел «социологический паспорт» со 125 классификациями), удалось дать столь точный результат.

Парадокс состоит в том, что при использовании квотных технологий *наличие многих априорных классификаций* — непреодолимая трудность, а *для изложенной методики — это благо*. Это обусловлено тем, что дисперсия итоговой оценки (5.9) имеет вид среднего гармонического дисперсий частных (неравноточных) измерений. Откуда следует: чем большее число вспомогательных номинальных шкал используется, тем меньше погрешность итогового результата (если, конечно, есть априорная статистика по этим классификациям).

Практика показала, что изложенный метод в реальных исследованиях (1991–2009 гг.) политологического, социологического и маркетингового характера обычно обеспечивает, при объемах случайного выборочного ансамбля 1500–2000 наблюдений, погрешности оценок $\hat{\nu}^k$ порядка 0.005–0.015.

6. Выборочные оценки частот встречаемости по социально-демографическим категориям населения

Изложенное выше стохастическое описание структурированной выборки позволяет решить одну из важных задач, которая *практически неразрешима в рамках традиционных для маркетинга и социологии «квотных» методов* эмпирических исследований. Эта задача — оценка частот встречаемости булевых признаков по социально-демографическим категориям населения.

На практике автором применялись различные методы для оценки частот встречаемости дихотомических признаков *по категориям населения* (Черепанов, 2006, 2007а, 2008). Ниже изложен один из наиболее простых методов оценивания этих частот, который, тем не менее, дает достаточно точные результаты.

Математическое ожидание условного распределения структурированной выборки 2-го рода (2.8) имеет вид

$$M[n_{i_j} - n_{ij}^k] = \frac{\theta_{ij}(1-\nu_{ij}^k)(n-n^k)}{1-\nu^k}. \qquad (6.1)$$

Это выражение позволяет записать оценку частоты встречаемости *k*-го признака по *j*-ой категории *i*-ой классификации в виде

$$\hat{\nu}_{ij}^{k} = 1 - \frac{1-\hat{\nu}^{k}}{\theta_{ij}} \frac{n_{ij} - n_{ij}^{k}}{n - n^{k}}. \qquad (6.2)$$

Состоятельность оценки (6.2) доказывается элементарно.

Дисперсия $\hat{\nu}_{ij}^{k}$ приближенно имеет вид

$$D\hat{\nu}_{ij}^{k} \cong \frac{(n_{ij} - n_{ij}^{k})\,(1-\hat{\nu}^{k})^{2}}{\theta_{ij}^{2}\,(n-n^{k})^{2}} \left(1 - \frac{n_{ij} - n_{ij}^{k}}{n - n^{k}} \right). \qquad (6.3)$$

Таким образом, в данной работе обоснован метод оценивания значений частот встречаемости булевых признаков по априорным классификациям (данные по которым есть в Росстате). Погрешности этих оценок сильно зависят от численности конкретной категории населения (точнее, ее доли в структуре населения). Но для объемов выборки n порядка 1.5–2.0 тыс. человек типичные значения погрешностей $3\sqrt{D\hat{\nu}_{ij}^{k}}$ составляют около 0.03–0.06.

7. Заключение

С момента возникновения массовых выборочных обследований до наших дней во всем мире тотально используются квотные методы получения и обработки данных. В статье показано, что использование квотной методологии с формальных вероятностно-статистических позиций корректно. Но при этом *использование квотных выборочных процедур сопряжено*:

- с невысокой точностью получаемых результатов для населения в целом;
- с невозможностью получить оценки частот встречаемости качественных признаков по социально-демографическим категориям;
- с высокой трудоемкостью формирования выборочного ансамбля, низкой оперативностью и высокой стоимостью получения данных.

Существует альтернатива квотной методологии выборочных исследований: работа со случайными выборками, при которой репрезентативность результатов обеспечивается на этапе математически корректной и достаточно нетривиальной машинной обработки данных обследований.

При использовании изложенных методов работы со случайными выборками точность оценок (по сравнению с «квотными» методами) *значительно возрастает,* стоимость опросов резко падает и оперативность исследований существенно повышается. *А возможность анализа общественного мнения в «разрезах» по категориям населения радикально повышает информативность экспертного анализа социальных и экономических проблем.*

Список литературы

Азаров С. В., Черепанов Е. В. (2004). Регрессионные методы статистического оценивания в социальных исследованиях. В кн.: *Математические методы и компьютерные технологии в маркетинговых и социальных исследованиях*. М.: Академия менеджмента инноваций, 2004, 56–72.

Азаров С. В., Пашин Ю. А., Черепанов Е. В. (2005). Современные компьютерные технологии в социальных исследованиях. *Безопасность Евразии*, 1, 264–281.

Айвазян С. А., Мхитарян В. С. (2001). *Прикладная статистика и основы эконометрики*. В 2-х томах. М.: Юнити.

Бернулли Я. (1986). *О законе больших чисел*. Пер. с лат. Юбилейное издание с предисловиями А. А. Маркова и А. Н. Колмогорова. М.: Наука.

Вучков И., Бояджиева А., Солаков Е. (1987). *Прикладной линейный регрессионный анализ*. М.: Финансы и статистика.

Кокрен У. (1976). *Методы выборочных исследований*. М.: Статистика.

Колочков Ю. М., Савелов В. И., Черепанов Е. В. (1990). Статистики бинарного отношения на множествах. В кн.: *Проблемы перспективного планирования и управления*. М.: изд. Госплана СССР, 88–98.

Косолапов М. С. (1997). Принципы построения многоступенчатой вероятностной выборки для субъектов Российской Федерации. *Социологические исследования*, 10, 98–109.

Миттаг Х.-Й., Ринне Х. (1995). *Статистические методы обеспечения качества*. М.: Машиностроение.

Мудров В. И., Кушко В. Л. (1976). *Методы обработки измерений*. М.: Советское радио.

Мхитарян В. С., Черепанов Е. В. (2006). Проблемы прикладной социологии в их привязке к социально-экономическим исследованиям. В кн.: *Информатика, социология, экономика, менеджмент.* Вып. 3, ч. 2. М.: Академия менеджмента инноваций, 23–33.

Орлов А. И. (2006). *Прикладная статистика.* М.: Экзамен.

Свешников А. А. (1972). *Основы теории ошибок*. Л.: ЛГУ.

Справочник по теории вероятностей и математической статистике. (1978). Под ред. В. С. Королюка. Киев: Наукова думка.

Толстова Ю. Н. (2000). *Анализ социологических данных*. М.: Научный мир.

Черепанов Е. В. (2006). *Вероятностно-статистические основы прикладной социологии и маркетинговых исследований.* М.: Академия менеджмента инноваций.

Черепанов Е. В. (2007а). *Статистическая методология для задач социологических и социально-экономических исследований*. М.: Академия менеджмента инноваций.

Черепанов Е. В. (2007б). К вопросу корректности использования стохастического формализма в социологических и социально-экономических исследованиях. *Безопасность Евразии*, 2 (28), 386–402.

Черепанов Е. В. (2007в). Негосударственное пенсионное страхование: состояние и перспективы (по результатам ряда социологических исследований 2006 года). *Социальная политика и социология*, 2 (34), 87–98.

Черепанов Е. В. (2007г). Социологический анализ структуры пользователей страховых услуг (на примере региональных исследований 2006 года по страхованию жизни и страхованию от несчастных случаев). *Социальная политика и социология*, 4 (36), 78–89.

Черепанов Е. В. (2007д). Стохастическое описание выборочного метода. *Социология: методология, методы, математическое моделирование*, 25, 167–189.

Черепанов Е. В. (2008). *Стохастические методы прикладной социологии и маркетинга рынков.* М.: Академия менеджмента инноваций.

Янке Е., Эмде Ф., Леш Ф. (1977). *Специальные функции*. М.: Наука.