Орехов А.В., Шебеко А.В.
Санкт-Петербургский государственный университет

Алгоритм многопараметрической корректировки выборочных данных

1. Введение. Основные результаты классической статистики основаны на очень сильном предположении, что генеральная совокупность числовых значений изучаемого параметра **A** имеет "хорошее" распределение, которое либо является равномерным, либо аппроксимируется одной из разновидностей нормального закона.

На самом же деле, это далеко не всегда так. Например, выборочное среднее наилучшим образом оценивает математическое ожидание нормально распределенной генеральной совокупности. Однако, средняя величина очень плохо "реагирует" на наличие выбросов (резко выделяющихся значений) в выборке. Их появление, если исключить грубые ошибки измерения, как раз и обусловлено нарушением "центрированности" распределения числовых значений параметра **A** в генеральной совокупности.

Еще "драматичнее" ситуация тогда, когда множество значений изучаемой случайной величины не имеет ни метрики, ни отношения порядка. Такое, например, сплошь и рядом случается при проведении социальных исследований.

Неправильное определение закона распределения в генеральной совокупности почти всегда приводит к фатальным ошибкам применения выборочного метода. Классическим примером его неудачного использования является следующая история [1].

2. Контрпример. Американский журнал "Литературное обозрение" в 1936 году провел социологический опрос. Целью исследования было предсказание итогов очередных выборов президента США, на пост которого в том году претендовали Ф.Д. Рузвельт и А.М. Ландон.

Отобрав случайным образом огромное количество адресов по телефонным книгам, редакция разослала по всей стране почтовые открытки с просьбой ответить на вопрос об отношении к кандидатам в президенты. Обработав ответы, которых было, кстати, гораздо меньше начального количества почтовых извещений, журнал объявил,

что с большим перевесом должен победить Ландон.

Однако вышло все наоборот – победил Франклин Делано Рузвельт. Следовательно, при проведении выборочного исследования была допущена ошибка. Ее суть заключалась в следующем. Отбор адресов респондентов из телефонных книг основывался на ложном предположении, что все избиратели США, как генеральная совокупность, имеют равномерное распределение предвыборных симпатий в различных социальных группах.

Ответ на вопрос об отношении к любому из кандидатов в президенты для каждого отдельного респондента имел распределение Бернулли (либо "за", либо "против"). Но совокупность таких ответов для различных социальных групп американского общества того времени имела свою специфику. Поэтому разные варианты ответа на вопрос об отношении к кандидатам в президенты не только имели неодинаковые вероятности, но и эти вероятности меняли свое значение в различных слоях американского общества. А именно, относительно рассматриваемого случая, в первой половине 20-го века на территории США личные телефоны имели более зажиточные американцы. Ситуацию усугубил тот факт, что привычку отвечать на письма имеют, в основном, представители деловых кругов. В результате получилось, что большую часть выборочной совокупности в данном исследовании составляли богатые американцы и особенно бизнесмены, интересы которых как раз и представлял А.М. Ландон. Ошибка при определении характера распределения в генеральной совокупности повлекла за собой нарушение принципа случайного отбора в выборку.

Описанное явление, когда выборка представляет не всю генеральную совокупность, а только какую-то ее часть, или в основном эту часть, называется смещением выборки. Таким образом, помимо грубых ошибок измерения, именно смещение выборочной совокупности приводит к возникновению ошибок при использовании выборочного метода.

Для преодоления указанной проблемы используется, например, послойное извлечение данных в выборочную совокупность. При этом слои в генеральной совокупности выбираются таким образом, чтобы значения параметров в каждом из них имели "хорошее" распределение. Или еще, для получения "правильной" выборки используется "квотирование" респондентов по различным социально-

демографическим характеристикам, распределение которых в генеральной совокупности известно, например, согласно результатам последней переписи населения.

Различные теоретические аспекты выборочных исследований, возникающие, в том числе, при послойном извлечении в выборочную совокупность или квотировании, наиболее полно изучены и описаны в широко известной монографии У. Кокрена [2].

3. Постановка задачи. Из сказанного выше следует, что необходимым условием достоверности результатов выборочного исследования является соответствие пропорций в выборке пропорциям в генеральной совокупности. Тогда выборочная совокупность достаточно хорошо воспроизводит генеральную совокупность во всех ее проявлениях. Следовательно, одной из основных проблем статистического анализа экспериментальных данных является получение несмещенных выборочных совокупностей, которые называются представительными или репрезентативными. Статистические оценки, вычисляемые как числовые характеристики выборочной совокупности, являются функциями результатов наблюдений. В соответствии с законом больших чисел они могут служить приближениями соответствующих характеристик генеральной совокупности. При этом точность приближения зависит, в том числе, и от "степени репрезентативности" (это понятие будет формально определено ниже) выборочной совокупности.

Вполне естественным является желание получить репрезентативную выборку вычислительными методами. Такая процедура приведения структуры выборочной совокупности в соответствие со структурой генеральной совокупности по одному или нескольким контролируемым параметрам называется "корректировкой выборки". Контролируемыми могут быть любые признаки, инвариантные задачам исследования, генеральные распределения которых известны. Например, в социальных исследованиях используют социально-демографическую информацию: распределения по полу, возрасту, образованию, семейному положению, типу местожительства и тому подобное.

Перейдем теперь к формальной стороне рассматриваемого вопроса. Пусть в случайном эксперименте изучается одновременно h параметров генеральной совокупности. Поставим в соответствие про-

извольному случайному эксперименту вектор $\mathbf{A} = (A_1, \ldots, A_h)$, где A_i – значение i-го параметра, $1 \leq i \leq h$. Не умаляя общности будем считать, что каждый параметр A_i принимает конечное число значений: $a_i^1, \ldots, a_i^{\beta_i}$.

Обозначим вектор $\mathbf{A}$ по-другому:

$$\mathbf{A} = (a_1^1, \ldots, a_1^{\beta_1}, \ldots, a_h^1, \ldots, a_h^{\beta_h}).$$

Здесь $\sum_{i=1}^{h} \beta_i = m$; $a_i^j = 1$, если значение параметра A_i равно a_i^j, и $a_i^j = 0$ – в противоположном случае. Ниже a_i^j будут называться признаками. С одной стороны, это, конечно, приведет к увеличению размерности рассматриваемых векторов до m, но с другой стороны, если рассматривать $\mathbf{A}$ как случайный вектор, то все его компоненты (признаки) будут иметь распределение Бернулли.

Пусть нам a priori известен закон распределения числовых значений некоторых параметров генеральной совокупности. Будем называть такие параметры контрольными. Не умаляя общности, будем считать, что это $A_1, \ldots, A_h$.

Рассмотрим $\{Y_k\}_{k=1}^N$ – генеральную совокупность, где $Y_k = (y_1, \ldots, y_m)$ – случайный m-мерный вектор. Каждая компонента y_i $(1 \leq i \leq m)$ имеет распределение Бернулли, т.е. Y_k – "нуль-один" случайный вектор. По прежнему будем называть i-ю компоненту случайного вектора i-м признаком.

Пусть $\mathbf{P} = (p_1, \ldots, p_m)$ – вектор вероятностей (генеральных долей) событий $y_i = 1$ и $\{X_s\}_{s=1}^n = \{X_1, \ldots, X_n\} = \{Y_{k_1}, \ldots, Y_{k_n}\}$ – выборка элементов из генеральной совокупности $\{Y_k\}_{k=1}^N$. Случайный вектор из выборочной совокупности имеет вид $X_s = (x_1, \ldots, x_m)$, где каждая компонента x_i, $1 \leq i \leq m$, также имеет распределение Бернулли.

Обозначим через $\mathbf{W} = (w_1, \ldots, w_m)$ вектор, компонентами которого являются относительные частоты (выборочные доли) значений $x_i = 1$.

Определение 1. Выборочная совокупность называется *репрезентативной (представительной) по контрольным параметрам* $A_1, \ldots, A_h$, если для любого i такого, что $1 \leq i \leq m$ будет выполняться равенство: $w_i = p_i$, т.е. $\mathbf{P} = \mathbf{W}$.

На практике этого добиться фактически невозможно, за исключением ряда частных случаев. В этой связи и возникает проблема

корректировки распределения выборочных долей контрольных параметров таким образом, чтобы для любого наперед заданного $\delta > 0$ выполнялось неравенство для евклидовой нормы $\| \mathbf{P} - \mathbf{W} \| < \delta$.

Определение 2. Будем называть число $\delta > \| \mathbf{P} - \mathbf{W} \|$ *степенью репрезентативности* выборочной совокупности. Чем меньше δ, тем выше степень репрезентативности выборочной совокупности.

Когда компоненты случайного вектора из выборочной совокупности подчиняются распределению Бернулли, возможна алгоритмическая корректировка выборки.

4. Формальное описание алгоритма. Итерации алгоритма корректировки выборочной совокупности схематично можно описать следующим образом.

Сначала выбирается признак, по которому разность между соответствующими выборочной и генеральной долями имеет максимальное значение. Если таких признаков несколько, то выбирается любой из них. Пусть это будет i-й признак.

Затем, в зависимости от знака разности $p_i - w_i$, либо из выборочной совокупности случайным образом удаляется некоторый вектор X_s, в котором компонента $x_i = 1$ (это происходит если $p_i - w_i < 0$), либо в выборочной совокупности случайным образом дублируется вектор X_s, в котором компонента $x_i = 1$, если $p_i - w_i > 0$.

Корректировка продолжается до тех пор, пока $\| \mathbf{P} - \mathbf{W} \|$ не станет меньше наперед заданного δ.

Обоснованием для применения этого алгоритма является тот факт, что когда $n \ll N$, гипергеометрическое распределение ничтожно мало отличается от биномиального, т.е. нет существенного различия между повторными и бесповторными выборками.

На каждом шаге мы, вообще говоря, будем получать новую выборочную совокупность, в которой компоненты вектора $\mathbf{W}$ будут принимать новые значения, отличные от предыдущих. Формирование новой выборки можно рассматривать как некоторое случайное событие $\mathbf{\Omega_k}$, которому, кстати, будет соответствовать определенное значение нормы, которое в дальнейшем будем обозначать $\| \mathbf{P} - \mathbf{W_k} \|$.

Рассмотрим последовательность: $\mathbf{\Omega_1}, \ldots, \mathbf{\Omega_k}, \ldots$. Этим случайным событиям можно поставить в соответствие двухэлементное множество исходов (состояний) $\{C, B\}$, где C – событие: $\| \mathbf{P} - \mathbf{W_k} \| \leq \| \mathbf{P} - \mathbf{W_{k+1}} \|$, т.е. "степень репрезентативности" выборочной

совокупности не увеличилась, так как не уменьшилась величина $\| \mathbf{P} - \mathbf{W} \|$, и где B – событие: $\| \mathbf{P} - \mathbf{W_k} \| > \| \mathbf{P} - \mathbf{W_{k+1}} \|$, т.е. "степень репрезентативности" выборочной совокупности увеличилась, а норма $\| \mathbf{P} - \mathbf{W} \|$ – уменьшилась.

Таким образом, $\mathbf{\Omega_1}, \ldots, \mathbf{\Omega_k}, \ldots$ – последовательность случайных событий, которая является цепью Маркова с двухэлементным множеством состояний $\{C, B\}$, так как вероятность наступления либо C, либо B зависит только от вида выборочной совокупности в данный момент времени, и не зависит от всех остальных предыдущих состояний. Эта цепь Маркова является неоднородной, с матрицей переходных вероятностей размера 2×2. При увеличении числа итераций все четыре вероятности переходов стремятся к $1/2$. Однако, подробное исследование сходимости этого алгоритма является отдельной, достаточно сложной теоретической задачей.

Литература

1. Венецкий И.Г., Венецкая В.И. Основные математико-статистические понятия и формулы в экономическом анализе: Справочник. М.: Статистика, 1979. 447 с.

2. Кокрен У. Методы выборочного исследования. М.: Статистика, 1976. 440 с.