П.С. Ростовцев
(Новосибирск)

ПЕРЕСТАНОВОЧНЫЙ КРИТЕРИЙ ДЛЯ АНАЛИЗА ВЗВЕШЕННОЙ ВЫБОРКИ[1]

Взвешивание данных в социологии применяется весьма часто. Вероятно, без этого не обойтись, но надежность статистического анализа взвешенных данных вызывает сомнение. Вся статистическая теория проверки гипотез становится бессильной. В работе рассматривается критерий сравнения средних значений переменных, основанный на перемешивании данных. Частными случаями этого критерия являются критерий Фишера для дихотомических переменных, критерий Вилкоксона для сравнения средних рангов. Оказалось, что в ряде случаев он дает возможность статистического анализа взвешенных данных.

Ключевые слова: выборка, взвешенная выборка, смешанная выборка, выборка без возвращения, перемешивание данных, сравнение средних, смещенные статистики, несмещенные статистики, статистический эксперимент.

Критерии, основанные на перемешивании данных, используются преимущественно в непараметрических методах. В известной книге Кендалла и Стьюарта [1, с. 622–686] перестановочный критерий упоминался при исследовании устойчивости параметрических методов к отклонениям распределения исследуемой совокупности от нормального. В данной работе он оказался полезным в силу активного применения статистического экспери-

Петр Симонович Ростовцев – кандидат технических наук, старший научный сотрудник Института экономики и организации промышленного производства СО РАН.

[1] Исследование поддержано грантом РФФИ 00-06-80221-а.

мента при анализе множественных сравнений в таблицах сопряженности неальтернативных вопросов [2, с. 148–164], а также и в методе детерминации моделей [3]. Классические параметрические статистики, используемые в этих методах, не соответствовали применяемой в них процедуре перемешивания. Это вызвало потребность в доработке математического аппарата.

Работа принесла пользу также и в частичном решении проблемы анализа взвешенных данных. Такие данные появляются в результате использования смещенных выборок. Приписывание веса объектам приводит к тому, что вычисление значимостей связей и смещений средних становится некорректным. Удается добиться лишь возможности получения несмещенных описательных статистик.

Модель перемешивания

Критерий Стьюдента для сравнения средних предполагает, что генеральная совокупность в условиях нулевой гипотезы нормальна (см., например, [3, с. 308–309]). Это конечно жесткое, трудновыполнимое для социально-экономических данных условие. Попытаемся исходить из другой модели, в которой мы, возможно, проигрываем в этой модельной ситуации, но выигрываем при анализе реальных данных.

Пусть $\{X_1, ..., X_N\}$ – совокупность наблюдений и среди них имеется группа наблюдений A. Рассматривается разбиение совокупности на *2* группы объектов A и $\overline{A}$, состоящие соответственно из N_A и $N_{\overline{A}}$ наблюдений ($N_A + N_{\overline{A}} = N$), и сравниваются средние $\overline{X}_A$ и $\overline{X}_{\overline{A}}$ в этих группах.

Гипотеза H_0 состоит в том, что наблюдения в группы A и $\overline{A}$ были отобраны случайно, и все $C_N^{N_A}$ вариантов такого отбора равновероятны. Альтернативной гипотезой является то, что этот отбор не случаен, а связан с различием средних.

Если A и $\overline{A}$ суть наличие и отсутствие некоторого свойства, то проверка нашей гипотезы является одновременно проверкой

независимости этого свойства и переменной X. Проверяя эту гипотезу, мы просто ограничиваемся конечной выборкой и фиксированными размерами отобранной из нее группы. При такой постановке задачи для проверки гипотезы нет необходимости полностью имитировать сбор данных, считать случайными размеры A и $\overline{A}$, считать распределение X подчиненным определенному закону X.

Вероятностная модель основана на предположении, что фиксированные наблюдения $\{X_1, \ldots, X_N\}$ перемешаны и случайно попали в подмножества A или $\overline{A}$.

Практически данный подход означает переход к анализу выборки без возвращения из конечной совокупности.

Перестановочный критерий для сравнения средних

Обычно генеральная совокупность исключительно велика и связь между объектами можно не принимать во внимание. Поэтому можно считать, что соблюдается условие независимости наблюдений, необходимое для центральной предельной теоремы.

В действительности все выборки в социально-экономических исследованиях производятся из конечной совокупности. Причем всегда эти выборки делаются без возвращения. Это означает, что элементы таких выборок взаимосвязаны. Некоторые работы учитывают эту особенность, например Г. Шварц [5] в своей работе вносит поправки в оценки дисперсии совокупности.

В данной работе схема выборки из конечной совокупности используется для исследования значимости отклонения средних. В основе лежит проверка, не случайно ли получено отклонение среднего (суммы) наблюдений по группе наблюдений. «Случайно» означает извлечение случайного подмножества наблюдений из конечного их множества.

Заметим, что

$$\overline{X}_A - \overline{X}_{\bar{A}} = \frac{N}{N_{\bar{A}}}(\overline{X}_A - \overline{X}), \tag{1}$$

где $\overline{X}$ – среднее по всей совокупности. Так как в соответствии с вероятностной моделью $\overline{X}$ – константа, исследование разности средних равносильно исследованию одного среднего $\overline{X}_A$ или даже суммы X для группы A.

Обозначим $Sum(X, A) = \sum_A X_i$.

Математическое ожидание Sum(X, A)

Пусть ξ_k – индикатор того, что наблюдение X_k попало в A, т.е. $\xi_k = 1$, если k-е наблюдение попало в A, и $\xi_k = 0$ в противном случае. Естественно, ξ_k имеет распределение Бернулли с параметром $p = N_A/N$. Так как $Sum(A)$ можно представить в виде

$$Sum(A) = \sum X_k \xi_k, \tag{2}$$

ясно, что математическое ожидание $Sum(A)$ имеет вид

$$E(Sum(X, A)) = N_A \overline{X}. \tag{3}$$

Дисперсия Sum(X, A)

Заметим, что

$$E(\xi_k \xi_m) = \frac{N_A}{N} \cdot \frac{N_A - 1}{N - 1} \tag{4}$$

Отсюда

$$E\left[(Sum(X, A))^2\right] = \frac{N_A N_{\bar{A}}}{N-1}\overline{X^2} + \frac{N_A - 1}{N-1} N_A N \overline{X} \tag{5}$$

Значит

$$D(Sum(X, A)) = N_A N_{\bar{A}} \frac{\overline{X^2} - \overline{X}^2}{N-1} \tag{6}$$

Таким образом,

$$D(Sum(X, A)) = \frac{N_A N_{\bar{A}}}{N} S^2(X), \tag{7}$$

где $S^2(X)$ вычисляется по формуле для обычной несмещенной оценки дисперсии; значит

$$D(\overline{X}_A) = \frac{N_{\bar{A}}}{N \cdot N_A} S^2(X) \tag{8}$$

Из этой формулы легко получить, что для дисперсии разности средних имеет место равенство

$$D(\overline{X}_A - \overline{X}_{\bar{A}}) = S^2(X)\left(\frac{1}{N_A} + \frac{1}{N_{\bar{A}}}\right) \tag{9}$$

Формула совершенно не отличается от аналогичной формулы для дисперсии разности средних в двух группах независимых наблюдений из одной бесконечной совокупности.

Введение индексных случайных величин позволяет также показать, что для двух непересекающихся случайных выборок из конечной совокупности A_1 и A_2, в сумме не составляющих $\{X_1, \ldots, X_N\}$, также верна формула

$$D(\overline{X}_{A_1} - \overline{X}_{A_2}) = S^2(X)\left(\frac{1}{N_{A_1}} + \frac{1}{N_{A_2}}\right) \tag{10}$$

Для двумерной совокупности наблюдений $\{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ аналогично можно получить, что

$$\operatorname{cov}(Sum(X, A), Sum(Y, A)) = \frac{N_A N_{\bar{A}}}{N} \operatorname{cov}(X, Y) \tag{11}$$

и

$$\operatorname{cov}(\overline{X}_A, \overline{Y}_A)) = \frac{N_{\bar{A}}}{N \cdot N_A} \operatorname{cov}(X, Y) \qquad (12)$$

Это означает, что в случае k-мерной конечной выборки $X_1, \ldots, X_N$ с ковариационной матрицей $\Sigma(X)$ ковариационная матрица $\overline{X}_A$ будет иметь вид

$$\sum(\overline{X}_A) = \frac{N_{\bar{A}}}{N \cdot N_A} \sum(X), \qquad (13)$$

а ковариационная матрица разности $\overline{X}_A - \overline{X}_{\bar{A}}$ –

$$\sum(\overline{X}_A - \overline{X}_{\bar{A}}) = \sum(X)\left(\frac{1}{N_A} + \frac{1}{N_{\bar{A}}}\right) \qquad (14)$$

Заметим, что в нашей вероятностной модели $E(Sum(X, A))$, $D(Sum(X, A))$, $\operatorname{cov}(Sum(X, A), Sum(Y, A))$ и, соответственно, $\Sigma(X)$ – не оценки, а точные значения математического ожидания, дисперсии и ковариации. То же можно сказать относительно дисперсий и ковариаций средних.

Статистики для перестановочного критерия

В качестве статистики критерия отклонения среднего мы возьмем величину

$$Z = \sqrt{N}\,\frac{Sum(X, A) - N_A \overline{X}}{\sqrt{N_A N_{\bar{A}}}\, S(X)}, \qquad (15)$$

где $Sum(A)$ – наблюдаемое значение суммы.

Первый момент статистики Z равен нулю, второй – единице. Третий момент выражается через третий выборочный момент μ_3 следующим образом:

$$E(Z^3) = \frac{N^{3/2}}{N_A^{1/2} \cdot N_{\bar{A}}^{1/2}} \cdot \frac{(N_{\bar{A}} - 1)}{(N-1)(N-2)} \cdot \frac{\mu_3}{S^3}. \qquad (16)$$

Если предположить, что соотношения между моментами также сходятся к некоторым разумным константам, то $E(Z^3)$ равно $O(1/N^{1/2})$, а значит, сходится к нулю. Можно показать также, что четвертый момент близок к 3. Таким образом, первые четыре момента Z близки к моментам стандартного нормального закона.

В книге Кендалла и Стьюарта [1] на основании исследования первых трех моментов перестановочной статистики $\omega = \frac{N_A}{N_{\bar{A}} S^2} (\overline{X}_A - \overline{X}_{\bar{A}})^2$, эквивалентной нашей статистике Z, утверждается, что статистика $t^2 = (N-2)\omega/(1-\omega)$ имеет распределение квадрата статистики Стьюдента с $N-2$ степенями свободы. Это значит, что и пятый и шестой моменты Z для большого числа наблюдений близки к моментам стандартного нормального закона.

Поскольку статистика Стьюдента имеет асимптотически нормальное распределение, значимость статистики Z целесообразно оценивать на основе нормального распределения.

Статистику критерия «перемешивания» для многомерных векторов имеет смысл брать в виде

$$R = \frac{N_A N_{\bar{A}}}{N} (\overline{X}_A - \overline{X}_{\bar{A}})^T \sum{}^{-1}(X)(\overline{X}_A - \overline{X}_{\bar{A}}) \qquad (17)$$

Статистика R инвариантна по отношению к невырожденным линейным преобразованиям X. Так как всегда переменные X можно преобразовать в некореллированные переменные Y, R представимо в виде суммы квадратов k некоррелированных, в асимптотике стандартных нормальных величин. Поэтому можно предположить, что эта статистика имеет в асимптотике распределение хи-квадрат с k степенями свободы.

Если ковариационная матрица вырождена (имеет ранг $r<k$), то, прежде чем вычислять статистику R, следует с помощью линейного преобразования, сохраняющего расстояние в пространстве X, перейти к пространству меньшей размерности.

Вычисление обратной ковариационной матрицы Σ^{-1} – дело сложное, тем более, когда требуется переход к пространству меньшей размерности. Вместо того, чтобы вычислять обратную матрицу, удобнее воспользоваться факторным анализом на основе метода главных компонент, которым снабжен практически любой статистический пакет. При этом должны использоваться все r компонент. Поиск главных компонент означает переход к ортогональному пространству с сохранением расстояния, вычисление факторов – стандартизацию главных компонент: обычно факторы – переменные с единичным стандартным отклонением и нулевым средним. Статистика R после применения факторного анализа примет вид

$$R = \frac{N \cdot N_A}{N_{\bar{A}}} \sum_{i=1}^{r} (\overline{F}_{i,A})^2, \qquad (18)$$

где $\overline{F}_{i,A}$ – среднее значение i-го фактора для группы A.

Частные случаи

Гипергеометрическое распределение. Это распределение представляет собой распределение N_{AB} в таблице сопряженности при независимости свойств A и B и фиксированных маргинальных частотах (см. рис. 1).

	B	$\overline{B}$	Итого
A	N_{AB}	$N_{A\bar{B}}$	$N_{A.}$
$\overline{A}$	$N_{\bar{A}B}$	$N_{A\bar{B}}$	$N_{\bar{A}.}$
Итого	$N_{.B}$	$N_{.\bar{B}}$	N

Рис. 1. Таблица сопряженности двух свойств

Пусть B_k – индикатор принадлежности k-го объекта группе B. Если положить $X_k = B_k$, получим $Sum(X, A) = N_{AB}$. Таким образом, гипергеометрическое распределение $N_{AB} = Sum(X, A)$ получается случайной выборкой подмножества A из $\{X_1, \ldots, X_N\}$.

Критерий Вилкоксона-Манна-Уитни. Пусть $\{X_1, \ldots, X_N\}$ – ранги какой-либо переменной. Тогда $Sum(A)$ – сумма рангов, которая, как известно, является критерием Вилкоксона для двухвыборочного теста на сдвиг распределения.

Легко убедиться, что довольно сложные формулы для вычисления дисперсий этих критериев, приведенные в учебниках (см., например, [3]), – частные случаи формулы (7). Таким образом, критерии Фишера и Вилкоксона и их нормальная аппроксимация – по сути один и тот же критерий.

Статистики для взвешенных данных

Пусть наблюдения имеют веса $\{\omega_1, \ldots, \omega_N\}$, с которыми вычисляются таблицы сопряженности и средних.

Взвешенные частоты

Допустим $X_k = \omega_k B_k$. Роль частоты N_{AB} пересечения A и B во взвешенной выборке играет сумма $Sum(X, A)$; распределение этой суммы здесь используется вместо гипергеометрического распределения. После того, как наблюдения X_k определены, получить Z-статистику не представляет сложности.

Следует заметить, что вес при этом, в условиях нулевой гипотезы, не связан с группой A, зато привязан к группе B.

Если рассмотреть клетки таблицы сопряженности (см. рис. 1), то для невзвешенной выборки Z-статистики для всех ее внутренних ячеек будут совпадать по абсолютной величине. Для взвешенной выборки в общем случае этого не наблюдается.

Взвешенное среднее в группе

Взвешенное среднее в группе A представляет собой выражение

$$\overline{X}_{A,\omega} = \frac{\sum_A \omega_k X_k}{\sum_A \omega_k} = \frac{\overline{\omega X}_A}{\overline{\omega}_A}, \qquad (19)$$

вычислять дисперсию и математическое ожидание которого в таблицах весьма сложно. Более удобной статистикой была бы, по аналогии с предыдущим, взвешенная сумма $Sum(\omega X, A) = \sum_A \omega_k X_k$, но, к сожалению, случайные изменения этой суммы связаны с весами объектов, так что даже если X – константа, существуют колебания $Sum(\omega X, A)$, обусловленные случайностью весов в группе A. Это легко увидеть, разложив эту сумму на два слагаемых

$$Sum(\omega X, A) = \sum_A \omega_k (X_k - c) + c \sum_A \omega_k, \qquad (20)$$

где $c = \frac{\overline{\omega X}}{\overline{\omega}}$ – взвешенное среднее по всей совокупности наблюдений.

Случайные отклонения первого слагаемого связаны с колебаниями X возле взвешенного среднего по всей совокупности, второго – с отклонениями суммы весов. Дисперсия суммы равна

$$D(Sum(\omega X, A)) = \frac{N_A N_{\bar{A}}}{N} (S^2(\omega(X - c)) + c^2 S^2(\omega) -$$

$$- 2c \operatorname{cov}(\omega(X - c), \omega)). \qquad (21)$$

Таким образом, с точностью до постоянного множителя, дисперсия взвешенной суммы по X состоит из взвешенной дисперсии отклонения X от взвешенного среднего ($S^2(\omega(X - c))$), из дис-

персии веса ($c^2S^2(\omega)$) и из удвоенной взвешенной ковариации отклонения и веса ($\mathrm{cov}(\omega(X-c), \omega)$) со знаком минус.

В случае невзвешенных данных ($\omega\equiv 1$) часть дисперсии, связанная с весом, равна нулю и дисперсия принимает обычный вид.

Z-статистика смещения взвешенной суммы

С содержательной точки зрения совершенно не имеет смысла рассматривать статистику отклонения среднего в группе A, имеющую самостоятельный источник дисперсии – вес объектов, вклад которого равен $c^2S^2(\omega)$. Поэтому целесообразно рассматривать центрированное значение X, равное $T = X - c$. Дисперсия T не будет включать такую составляющую, поскольку $\overline{\omega T} = 0$. В этом случае Z отклонение T от ожидаемого нулевого значения будет иметь вид

$$Z = Sum(\omega T, A) \Big/ \left(\sqrt{\frac{N_A N_{\bar{A}}}{N}} S(\omega T) \right) = \overline{\omega T}_A \Big/ \left(\sqrt{\frac{N_{\bar{A}}}{N \cdot N_A}} S(\omega T) \right) \quad (22)$$

Поскольку невзвешенные данные можно считать взвешенными с весом $\omega = 1$, то последняя формула дает Z-статистику для всех рассматриваемых случаев.

Статистика Z полезна тем, что отражает смещение значений группы A от взвешенного среднего значения на всей совокупности.

Z-статистика смещения взвешенной суммы векторов

Для анализа взвешенных данных целесообразно элиминировать влияние смещения веса, перейдя от вектора X к вектору $T = X - c$ так, чтобы $\overline{\omega T} = 0$. Здесь вектор c – обычное взвешенное среднее. Соответствующая статистика хи-квадрат будет иметь вид

$$R = \frac{N_A N_{\bar{A}}}{N} (\overline{\omega T}_A - \overline{\omega T}_{\bar{A}})^T \sum{}^{-1} (\omega T)(\overline{\omega T}_A - \overline{\omega T}_{\bar{A}}). \quad (23)$$

Взвешенный критерий хи-квадрат

Проверка сходства распределений номинальной переменной в двух группах объектов в обычных условиях проводится с помощью критерия хи-квадрат. По существу этот критерий – перестановочный. Это его свойство используется в точных статистических тестах в статистическом пакете SPSS [6]. Номинальную переменную можно представить совокупностью дихотомических переменных и далее анализировать отклонение средних по этому вектору, учитывая вырожденность ковариационной матрицы. Вычисленная по формуле (17) статистика хи-квадрат будет совпадать с классической.

Для анализа взвешенных данных следует эти переменные умножить на весовую переменную и отцентрировать по формуле, используемой в одномерном случае. Далее происходит сравнение средних по полученному вектору.

Следует заметить, что, поскольку сумма весов в группе здесь будет меняться, число степеней свободы у статистики хи-квадрат будет, вообще говоря, равно числу значений номинальной переменной.

Статистический эксперимент

Для демонстрации нормальности распределения теста в качестве данных взяты 100 значений переменной X, равные первым 100 натуральным числам: $X_1 = 1, \ldots, X_{100} = 100$, веса этих «наблюдений» $\omega_1 = \ldots = \omega_{50} = 0{,}75,\ \omega_{51} = \ldots = \omega_{100i} = 1{,}5,\ i = 51, \ldots, 100$. Заметим, что сумма весов здесь не равна числу объектов, но это и неважно.

Индикаторы группы A, переменные $I_1, \ldots, I_{100}$ генерировались случайно (перемешивались), так что 40 из них равны единице, 60 – нулю. Было проведено 5000 экспериментов по перемешиванию данных (генерации $I_1, \ldots, I_{100}$).

Несмотря на большое число экспериментов по перемешиванию данных (5000) и относительно небольшой объем данных (100 наблюдений), тест Колмогорова-Смирнова показал незначимое от-

личие распределения *Z*-статистики от стандартного нормального (наблюдаемый уровень значимости равен 0,4368). Результаты были получены с помощью пакета SPSS и иллюстрируются рис. 2 и 3.

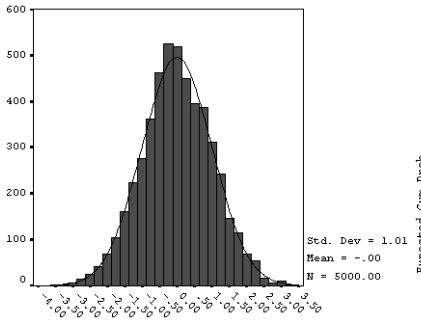

Рис. 2. Гистограмма распределения Z- статистики для взвешенных данных

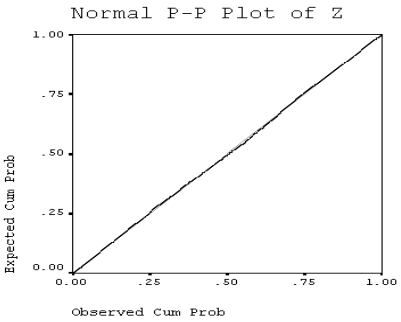

Рис. 3. Сравнение эмпирической и теоретической функций распределения Z. График «наблюдаемая частота-вероятность»

Сравнение формальных вычислений по взвешенным данным с перестановочными критериями

Для иллюстрации взяты данные анкетного опроса 1992 г., проведенного ИЭиОПП СО РАН, «Результаты реформы». Выборка была смещенной, она выравнивалась по полу, возрасту и образованию, причем весовые переменные определялись на основе государственной статистики. Для анализа использовались индикаторные переменные жизненных ценностей респондентов, возраст, логарифм душевого дохода, семейное положение, пол.

Формальное применение t-теста

Если работа интересна, то и заработок должен быть больше, да и душевой доход в семье выше. Как показывает таблица средних, средний логарифм душевых доходов несколько выше в группе респондентов, указавших в качестве жизненной ценности ин-

тересную работу, чем у остальных респондентов (по взвешенным данным и по невзвешенным).

Таблица 1

СРЕДНИЕ И ВЗВЕШЕННЫЕ СРЕДНИЕ ЛОГАРИФМА ДУШЕВОГО ДОХОДА

Интересная работа – жизненная ценность	Число наблюдений	Сумма весов	Среднее	Стандартное отклонение	Взвешенное среднее	Взвешенное стандартное отклонение
Да	393	349,4	8,34	0,81	8,375	0,781
Нет	664	704,4	8,20	0,83	8,274	0,715

Таблица 2

ЗНАЧИМОСТЬ РАЗЛИЧИЯ ГРУПП: Т-СТАТИСТИКА НА НЕВЗВЕШЕННЫХ ДАННЫХ, ФОРМАЛЬНОЕ ПРИМЕНЕНИЕ Т-СТАТИСТИКИ НА ВЗВЕШЕННЫХ ДАННЫХ И ПЕРЕСТАНОВОЧНЫЙ КРИТЕРИЙ

Статистика	Невзвешенные данные	Взвешенные данные	
		Формальное применение *t*-статистики	Перестановочный критерий
Разность средних	0,152	0,101	0,101
Статистика значимости	$t = 2,91$	$t = 2,09$	$z = 1,83$
Двусторонняя значимость различия средних	0,0037	0,0364	0,0673

В соответствии с таблицей по невзвешенным данным разница средних существенна (наблюдаемая значимость равна 0,0037), но этим результатам нельзя доверять, так как выборка скошена. Формальное вычисление по взвешенным данным дало наблюдаемый уровень значимости 0,0364, но этому результату также нельзя доверять, так как вычисление критерия некорректно. Перестановочный критерий, учитывающий реальные объемы групп и веса, дал значение *Z*-статистики, равное 1,83, и наблюдаемый уровень значимости 0,0673. Так что существенность различий групп здесь под сомнением.

Взвешенные частоты

Женатые люди чаще думают о семье, чем об интересной работе. С другой стороны, работа – это источник их существования. С целью иллюстрации метода мы попробовали изучить, большей или меньшей жизненной ценностью считают интересную работу семейные?

Табл. 3 склоняет нас поверить версии, что женатые реже, чем остальные респонденты, относят интересную работу к жизненным ценностям. Причем это прослеживается как по взвешенным данным, так и по невзвешенным (разница в долях указавших эту жизненную ценность составляет ~1–2%).

Таблица 3

ОТНОШЕНИЕ СЕМЕЙНЫХ К ИНТЕРЕСНОЙ РАБОТЕ

Интересная работа – жизненная ценность	Невзвешенные данные			Взвешенные данные		
	Женатые/ замужние	Не женат/ не замужем	Всего	Женатые/ замужние	Не женат/ не замужем	Всего
Да	345	111	456	290	122	412
	75,7	24,3	100,0	70,4	29,6	100,0
	38,4	35,8	37,7	34,3	33,5	34,1
Нет	554	199	753	555	242	797
	73,6	26,4	100,0	69,6	30,4	100,0
	61,6	64,2	62,3	65,7	66,5	65,9
Всего	899	310	1209	845	364	1209
	74,4	25,6	100,0	69,9	30,1	100,0
	100	100	100	100,0	100,0	100,0

При корректном расчете значимости отклонения в данном случае естественно «привязать» вес к дихотомии «семейные/несемейные». Здесь мы видим, что для взвешенных данных большее отклонение суммы весов в ячейке «Женат – работа ценность» от ожидаемого можно получить случайно с вероятностью 0,0214, значительно меньшей, чем 0,843, которую нам «обещает» формальное использование критерия Фишера (хотя, конечно, ясно, что использование его в данном случае выборки неверно). Таким образом, связь здесь обнаруживается на уровне значимости 0,02,

что более надежно, чем предполагалось при обычной некорректной процедуре.

Возникает вопрос, как же так, смещение всего на полпроцента обнаруживает значимую связь, в то время как для исходных данных смещения более чем на один процент – незначимы? Дело, по-видимому, в том, что вес объектов столбца «Женатые…» в основном несколько меньше единицы, поэтому и дисперсия переменной *ωB* несколько меньше, чем можно было бы ожидать, а значит, и небольшие изменения суммы весов в ячейке существенны.

Таблица 4

ДВУСТОРОННЯЯ ЗНАЧИМОСТЬ СМЕЩЕНИЯ ЧАСТОТ: ТОЧНЫЙ КРИТЕРИЙ ФИШЕРА ДЛЯ НЕВЗВЕШЕННЫХ ДАННЫХ, ФОРМАЛЬНОЕ ПРИМЕНЕНИЕ ЕГО ДЛЯ ВЗВЕШЕННЫХ ДАННЫХ И ПЕРЕСТАНОВОЧНЫЙ КРИТЕРИЙ

Критерий	Невзвешенные данные	Взвешенные данные
Формальное применение точного критерия Фишера	.455	.843
Перестановочный критерий	.455	0,0214

В подтверждение сказанного стоит обратить внимание на средние значения взвешенного индикатора группы семейных *ωB* в группах по оценке работы как жизненной ценности (табл. 5). Учитывая, что в нашей вероятностной модели стандартное отклонение среднего в группе равно $\sqrt{\frac{753}{1209 \cdot 453}}0{,}7362 = 0{,}0273$, различие средних достаточно велико.

Таблица 5

СРАВНЕНИЕ СРЕДНИХ ЗНАЧЕНИЙ ПЕРЕМЕННОЙ *ωB*

Интересная работа – жизненная ценность	Среднее *ωB*	N	Стандартное отклонение
Да	0,6370	456	0,5728
Нет	0,7375	753	0,8175
Всего	0,6996	1209	0,7362

Сравнение распределений по номинальной переменной. Взвешенный хи-квадрат

Не будем далеко уходить от темы предыдущего примера, сравним распределения по семейному положению в целом в тех же группах: группе считающих интересную работу жизненной ценностью и группе остальных респондентов. При этом, для сравнения результатов:

– проведем расчеты критерия хи-квадрат на невзвешенных данных, вычислив критерий хи-квадрат традиционным способом;

– покажем на примере, не приводя сложных доказательств, эквивалентность традиционного критерия хи-квадрат с полученным значением, используя факторный анализ для невзвешенных данных;

– вычислим наблюдаемую значимость различия распределений, формально применив критерий хи-квадрат для взвешенных данных;

– вычислим наблюдаемую значимость различия распределений с учетом взвешенности данных.

Итак, в табл. 6 приведены частотные распределения и процентные распределения. На первый взгляд различие распределений существенно.

Некоторое различие распределений прослеживается, особенно в процентах по вертикали. Значение критерия для таблицы хи-квадрат равно 8,295837 при наблюдаемой значимости 0,040277.

Для вторичного вычисления этой же статистики с помощью факторного анализа на вход этой процедуры мы задали переменные – индикаторы групп по семейному положению. Как видно из табл. 7, число компонент соответствует числу значений исследуемой номинальной переменной без единицы (что естественно).

На основании формулы (18) и таблицы средних для факторов, значение критерия R равно 8,288975, т.е. отличается в сотых долях от значения хи-квадрат 8,295837, полученного традиционным способом. Это отличие объясняется просто: при традиционном

вычислении хи-квадрата неявным образом используются смещенные оценки соответствующих переменных. Коэффициент смещения равен $N/(N-1)$, в частности, в нашем случае $N = 1209$ и оказывается, что 8,288975×1209/1208 = 8,295837 (!). Так что традиционный способ дает немного более оптимистичные выводы относительно связи семейного положения и интереса к работе. Наблюдаемый уровень значимости здесь оказался равным 0,040402.

Таблица 6

СЕМЕЙНОЕ ПОЛОЖЕНИЕ И ОТНОШЕНИЕ К ИНТЕРЕСНОЙ РАБОТЕ. НЕВЗВЕШЕННЫЕ ДАННЫЕ

(частоты, проценты по горизонтали, проценты по вертикали)

Интересная работа – жизненная ценность	Семейное положение				
	Женат/ замужем	Разведен(а)	Вдовец/вдова	Холост/ не замужем	Всего
Да	345 75,7 38,4	40 8,8 44,0	24 5,3 25,0	47 10,3 38,2	456 100 37,7
Нет	554 73,6 61,6	51 6,8 56,0	72 9,6 75,0	76 10,1 61,8	753 100 62,3
Всего	899 74,4 100	91 7,5 100	96 7,9 100	123 10,2 100	1209 100 100

Таблица 7

ДИСПЕРСИЯ, ОБЪЯСНЕННАЯ ГЛАВНЫМИ КОМПОНЕНТАМИ

(невзвешанные данные)

Компонента	Объясненная дисперсия	Процент дисперсии	Накопленный процент
1	1,815	45,371	45,371
2	1,102	27,538	72,910
3	1,084	27,090	100
4	0,000	0,000	100

Таблица 8

СРЕДНИЕ ЗНАЧЕНИЯ ФАКТОРОВ (невзвешенные данные)

Интересная работа – жизненная ценность	Факторы	F_1	F_2	F_3
Да	Среднее	–0,0295	–0,0352	0,0960
	N	456	456	456
	Стандартное отклонение	0,9839	0,9702	0,9706
Нет	Среднее	0,0179	0,0213	–0,0581
	N	753	753	753
	Стандартное отклонение	1,0098	1,0177	1,0136
Всего	Среднее	0	0	0
	N	1209	1209	1209
	Стандартное отклонение	1	1	1

Табл. 9 для взвешенных данных показывает те же смещения в распределении, может быть немного различающиеся количественно. Формально вычисленный критерий хи-квадрат равен 11,12592, что формально соответствует наблюдаемому уровню значимости 0,011064.

Таблица 9

СЕМЕЙНОЕ ПОЛОЖЕНИЕ И ОТНОШЕНИЕ К ИНТЕРЕСНОЙ РАБОТЕ. ВЗВЕШЕННЫЕ ДАННЫЕ

Интересная работа – жизненная ценность	Семейное положение				
	Женат/ замужем	Разведен(а)	Вдовец/вдова	Холост/ не замужем	Всего
Да	290 70,4 34,3	36 8,7 40,0	24 5,8 21,2	62 15,0 38,5	412 100 34,1
Нет	555 69,6 65,7	54 6,8 60,0	89 11,2 78,8	99 12,4 61,5	797 100 65,9
Всего	845 69,9 100	90 7,4 100	113 9,3 100	161 13,3 100	1209 100 100

Что же дала процедура вычисления критерия для взвешенных данных? Для вычисления используются индикаторные переменные групп по семейному положению, умноженные на весовую переменную. Здесь уже число полученных факторов совпадает с числом значений.

Таблица 10

ДИСПЕРСИЯ, ОБЪЯСНЕННАЯ ГЛАВНЫМИ КОМПОНЕНТАМИ (взвешанные данные)

Компонента	Объясненная дисперсия	Процент дисперсии	Накопленный процент
1	1,348	33,698	33,698
2	1,068	26,710	60,408
3	1,044	26,097	86,505
4	0,540	13,495	100

Для вычисления критерия используется таблица средних с четырьмя факторами (см. табл. 11).

Таблица 11

СРЕДНИЕ ЗНАЧЕНИЯ ФАКТОРОВ (взвешенные данные)

Интересная работа – жизненная ценность	Факторы	F_1	F_2	F_3	F_4
Да	Среднее	0,0421	–0,0086	–0,0776	–0,1055
	N	456	456	456	456
	Стандартное отклонение	0,8725	0,9707	0,7571	0,8100
Нет	Среднее	–0,0255	0,0052	0,0470	0,0639
	N	753	753	753	753
	Стандартное отклонение	1,0696	1,0179	1,1197	1,0948
Всего	Среднее	0	0	0	0
	N	1209	1209	1209	1209
	Стандартное отклонение	1	1	1	1

Значение критерия *R* здесь равно 13,75618, а наблюдаемый уровень значимости, вычисленный на основе распределения хи-

квадрата с 4 степенями свободы, равен 0,008115. Так что в данном случае корректная оценка обнаружила более значимую связь.

Практическое применение метода

Статистические пакеты обычно имеют средства вычисления средних, реализации факторного анализа на основе метода главных компонент, поэтому корректная работа с взвешенными данными вполне возможна. В Приложении мы привели программу в синтаксисе SPSS, которая проводит практически все необходимые расчеты и может быть изменена читателем для своих целей. Лишь небольшую часть расчетов целесообразно сделать с помощью Excel. Имеется в виду окончательное вычисление статистик на основе полученных средних и определение их значимости с использованием функции НОРМСТРАСП и ХИ2РАСП.

Непосредственное применение описанного метода нами реализовано в последнем варианте программы при получении таблиц для неальтернативных вопросов Typology Tables. В ней статистики перестановочного критерия даны для сравнения средних и смещений частот. Основным в работе является решение проблемы множественных сравнений при исследовании связи между неальтернативными вопросами.

В ближайшее время предполагается также сделать корректным анализ взвешенных данных в программе DAMO, осуществляющей множественное сравнение и детерминацию статистических моделей, соответствующих группам объектов.

Описание программ можно найти в http://ieie.nsc.ru:8101/~rokos/.

ЛИТЕРАТУРА

1. *Кендалл М.Дж., Стьюарт А.* Статистические выводы и связи. М:. Наука, 1973.

2. *Ростовцев П.С., Костин В.С., Олех А.Л.* Множественные сравнения в таблицах для неальтернативных вопросов // Анализ и моделирование экономических процессов переходного периода в России. Новосибирск, ИЭиОПП СО РАН, 1999. Вып. 4.

3. *Айвазян С.А., Мхиторян В.С.* Прикладная статистика и основы эконометрики. М.: Издательское объединение «Юнити», 1998.

4. *Ростовцев П.С., Костин В.С., Олех А.Л., Жданов А.С.* Автоматизация анализа социально-экономических данных: Детерминация моделей // Вестник новосибирского государственного университета: социально-экономические науки. 2000. № 1. С. 20–37.

5. *Шварц Г.* Выборочный метод. М.: Статистика, 1978.

6. SPSS: Exact tests 6.1 for windows. Chicago, 1995.

ПРИЛОЖЕНИЕ

Программы сравнения взвешенных средних и распределений в синтаксисе SPSS

```
* ПЕРЕМЕННЫЕ.
* Wespvo – вес, построенный по полу, возрасту,
* образованию.
* InWork – индикатор жизненной ценности «Интересная
* работа».
* Fstatus – «Семейное положение».
* Income – «Душевой доход».
compute Married = (Fstatus=1).
compute Devorce = (Fstatus=2).
compute Widow = (Fstatus=3).
compute Single = (Fstatus=4).
Variable labels Married «Женат/замужем» Devorce
«Разведен(а)»
Widow «Вдова/вдовец» Single «Холост/не замужем».
Value labels Married Devorce Widow  Single 1 «Да» 0
«Нет».
* ———————————————————.
* ЦЕННОСТЬ «ИНТЕРЕСНАЯ РАБОТА» И СМЕЩЕНИЕ СРЕДНЕГО
* ЛОГАРИФМА ДУШЕВОГО ДОХОДА.
Compute  lincome=ln(income).
Weight by Wespvo.
DESCRIPTIVES  VARIABLES=lincome /STATISTICS=MEAN .
* среднее логарифма доходов (8,307743) используется для
* построения переменной wlincome.
compute wT=Wespvo*(lincome - 8.307743).
Weight off.
```

```
MEANS   TABLES=wT  BY InWork   /CELLS MEAN COUNT STDDEV.
* получены: N=1057, NA=393 (с учетом неопределенных
* кодов), S=0,821434, среднее значение wTA = 0,060067, по
* формуле (22) Z=1,83.
* ————————————————————————————.
* ЦЕНЯТ СЕМЕЙНЫЕ ИНТЕРЕСНУЮ РАБОТУ?
compute wB=Married *wespvo.
Weight off.
MEANS  TABLES= wB BY InWork   /CELLS MEAN COUNT STDDEV.
* среднее значение в целом по совокупности  wB равно
* 0,093373, в группе равно 0,052621, при объеме группы
* NA=456 и всей совокупности N=1209.
* Отсюда Z-статистика равна -2,30.
* ————————————————————————————.
* СВЯЗЬ ЦЕННОСТИ «ИНТЕРЕСНАЯ РАБОТА» С СЕМЕЙНЫМ
* ПОЛОЖЕНИЕМ.
compute WMarried = Married*Wespvo.
compute WDevorce = Devorce*Wespvo.
compute WWidow = Widow*Wespvo.
compute WSingle = Single*Wespvo.
Weight off.
FACTOR
  /VARIABLES wmarried wdevorce wwidow wsingle  /MISSING
LISTWISE /ANALYSIS
  wmarried wdevorce wwidow wsingle
  /PRINT INITIAL EXTRACTION  /CRITERIA FACTORS(4)
ITERATE(25)
 /EXTRACTION PC  /ROTATION NOROTATE  /SAVE REG(ALL)  /
METHOD=CORRELATION .
MEANS   TABLES=fac1_1 fac2_1 fac3_1 fac4_1  BY inwork  /
CELLS MEAN COUNT  .
* Сумма квадратов средних значений факторов, домноженная
* на N*NA/(N-NA) - искомая статистика хи-квадрат (R),
* равна 13,75618.
```