СТАТИСТИЧЕСКИЙ ЭКСПЕРИМЕНТ

В.С. Костин, С.А. Ермаханова
(*Новосибирск*)

СТАТИСТИЧЕСКИЙ ЭКСПЕРИМЕНТ ДЛЯ ПРОВЕРКИ АДЕКВАТНОСТИ РЕЗУЛЬТАТОВ АНАЛИЗА ПАРНЫХ СВЯЗЕЙ

В статье предложена методика проверки корректности использования различных способов анализа связи между двумя переменными. Она опирается на полный (сплошной) поиск связей по матрице данных типа «объект-признак» и включает серию статистических экспериментов с перемешиванием данных. Ее возможности иллюстрируются на примере анализа результатов экспертного опроса.

Ключевые слова: меры связи, дисперсионный анализ, хи-квадрат, гипергеометрическое распределение, значимость связи, распределение значимости, статистический эксперимент, критерий согласия.

Постановка исследовательской задачи

Традиционно анализ связей в социологических исследованиях сводится к применению статистических критериев для проверки гипотез о наличии парных связей для небольшого числа сочетаний переменных. Пары переменных для проверки гипотез выбираются исследователем из содержательных соображений. Применение методов математической статистики позволяет социоло-

Виталий Сергеевич Костин – старший научный сотрудник Института экономики и организации промышленного производства СО РАН. E-mail: kostin@ieie.nsc.ru.
Салтанат Амангелдыкызы Ермаханова – кандидат социологических наук, младший научный сотрудник Института экономики и организации промышленного производства СО РАН. E-mail: essaltanat@mail.ru.

гу достаточно объективно судить о наличии связей в данных. Тем не менее, такой подход все же оставляет без внимания некоторые источники ошибок в результатах. В этой связи мы предлагаем методические приемы дополнительной проверки корректности проведенного анализа и обоснованности выводов. Следует подчеркнуть, что эти приемы требуют программного обеспечения для автоматизации многократного выполнения статистических расчетов.

Первый прием связан с проверкой корректности применения выбранной меры связи. Контроль при этом сводится к сравнению теоретического распределения статистики, по которой судят о наличии связи, с эмпирическим распределением той же статистики, полученным в случайных экспериментах с перемешиванием данных, которые адекватно воспроизводят условия нулевой гипотезы о независимости переменных. Если гипотеза *о согласованности* этих распределений отвергается, исследователь вправе сделать вывод о некорректности применения меры связи для выбранной пары переменных для случая исследуемой выборки. Тогда следует использовать другую меру связи либо вовсе отказаться от проверки наличия связи между этой парой переменных.

Первый прием позволяет убедиться в возможности проверки каждой конкретной гипотезы выбранным способом. Как известно, любой статистический критерий явно или неявно требует выполнения некоторых предположений на исходных для анализа данных, которые не всегда легко проверить. Но он может удовлетворительно работать и при определенных нарушениях таких предположений. С одной стороны, перестраховываясь, подходя слишком строго к контролю этих предположений, исследователь во многих случаях вынужден отказаться от проверки гипотез. С другой стороны, подходя слишком мягко, мы рискуем получить необоснованные или даже ошибочные выводы о наличии или отсутствии связи. Предлагаемый нами подход на основе дополнительных вычислений позволяет контролировать корректность статистического критерия для каждой пары переменных. Тем самым появляет-

ся инструмент достаточно гибкого отсева гипотез, не подлежащих корректной проверке.

Второй прием позволяет проверить гипотезу об отсутствии *значимых парных связей* в матрице данных. Он основан на том, что совокупное распределение значимостей для всех прошедших через фильтр гипотез при наличии связей должно отклоняться от равномерного в сторону преобладания значимостей, близких к нулю. Если это не наблюдается, то появляются основания для вывода, что исходная для анализа матрица данных не содержит информации о парных связях. В таком случае, скорее всего, данные следует признать непригодными для дальнейшего исследования как недостаточно информативные. Тем самым появляется возможность оценки качества отдельного эмпирического исследования.

В математической статистике разработаны различные способы анализа связи. В случае двух переменных они зависят от уровня их измерения. Но все они исходят из двух предположений. Во-первых, прежде чем анализировать связь между переменными, необходимо убедиться в ее наличии. Во-вторых, проще проверить отсутствие связи, чем ее наличие, поскольку связь может проявляться во множестве различных форм, а ее отсутствие – в единственной форме. Для проверки отсутствия связи формулируют так называемую нулевую гипотезу о том, что две рассматриваемые переменные являются независимыми. Как известно, независимость случайных величин можно выразить строго через вероятность наблюдения совместного события [1, с. 382]:

$$P(A_iB_j) = P(A_i)\,P(B_j) \qquad (1)$$

для всех A_i и B_j. Другими словами, при независимости признаков значение, «принятое» признаком A, не влияет на вероятности возможных значений признака B и наоборот. В этом случае условные вероятности событий A_i и B_j равны безусловным вероятностям:

$$P(B_j|A_i) = P(A_iB_j) / P(A_i) = P(B_j) \qquad (2)$$

$$P(A_i|B_j) = P(A_iB_j) / P(B_j) = P(A_i) \qquad (3)$$

Здесь $P(A_i)$, $P(B_j)$ – безусловные вероятности событий A_i и B_j, $P(A_iB_j)$ – вероятность совместного события A_i и B_j, $P(B_j|A_i)$ – вероятность события B_j при условии наступления события A_i, $P(A_i|B_j)$ – аналогично, вероятность события A_i при условии B_j. Из этих соотношений следует, что при *случайном перемешивании* (изменении порядка следования значений в массиве данных) для любой из двух переменных (или обеих) в точности выполняется предположение нулевой гипотезы – переменные становятся независимыми, что мы и будем использовать в дальнейшем.

Прежде чем рассмотреть результаты статистического эксперимента, коротко остановимся на специфике известных способов анализа связи переменных, каждая из которых может иметь собственный уровень измерения.

Связь между двумя номинальными переменными

Традиционно связь между номинальными переменными оценивается на основе статистики *хи-квадрат*. Рассмотрим таблицу сопряженности[1], построенную по переменным v2 и v67 (см. табл. 1), в клетках которой последовательно представлены три значения:

n_{ij} – наблюдаемые частоты, e_{ij} – ожидаемые частоты, z_{ij} – стандартизованные отклонения.

Для дальнейшего изложения введем и другие обозначения: N – объем выборки, а $n_{i\circ} = \sum_{j=1}^{c} n_{ij}$ и $n_{\circ j} = \sum_{i=1}^{r} n_{ij}$ – маргинальные час-

[1] Приводится пример из исследования, носящего характер экспертного опроса. Оно проводилось в июне-августе 2006 г. отделом социальных проблем ИЭОПП СО РАН. Экспертами являлись высококвалифицированные управленцы высшего звена и профессионально-компетентные специалисты высокого ранга, занятые в разных сферах: государственная служба; торгово-промышленный бизнес; наука, культура и высшее образование; социальная работа. Эксперты на момент опроса проживали и работали в следующих городах: Алматы, Астана, Жезказган, Усть-Каменогорск, Семей, Тараз. Объем выборки – 260 человек.

тоты соответственно по строкам и по столбцам, r – число строк, c – число столбцов.

Таблица 1

НАБЛЮДАЕМЫЕ И ОЖИДАЕМЫЕ ЧАСТОТЫ, СТАНДАРТИЗОВАННЫЕ ОТКЛОНЕНИЯ

V67: Хотите ли Вы, чтобы влияние внешней культуры на казахстанскую молодежь усиливалось?	V2: Пол		Итого
	Мужской	Женский	
Да	38 45,70 –1,96	71 63,30 1,96	109
Нет	71 63,30 1,96	80 87,70 –1,96	151
Итого	109	151	260

В предположении независимости ожидаемые частоты равны: $e_{ij} = P(A_i)P(B_j)N = \frac{n_{i\circ}}{N} \cdot \frac{n_{\circ j}}{N} N = \frac{n_{i\circ} \cdot n_{\circ j}}{N}$. Если нулевая гипотеза верна, то наблюдаемые частоты должны быть достаточно близки к ожидаемым. Для оценки близости вводится статистика *хи-квадрат*, вычисляемая как взвешенная сумма квадратов отклонений наблюдаемых частот от ожидаемых [1, с. 787–789]:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}. \qquad (4)$$

Нулевую гипотезу можно сформулировать, как равенство нулю этой статистики. В нашем случае:

$$\chi^2 = \frac{(38 - 45{,}7)^2}{45{,}7} + \frac{(71 - 63{,}3)^2}{63{,}3} + \frac{(71 - 63{,}3)^2}{63{,}3} + \frac{(80 - 87{,}7)^2}{87{,}7} =$$

$$= 1{,}29 + 0{,}94 + 0{,}94 + 0{,}67 = 3{,}84 \qquad (5)$$

На рис. 1 по горизонтальной оси отложено значение статистики *хи-квадрат*, по вертикальной оси – вероятность случайно получить большее или равное значение этой статистики в условиях нулевой гипотезы. Видно, что с ростом значения статистики вероятность ее наблюдения убывает достаточно быстро, т.е. основная масса реализаций сосредоточена в непосредственной близости от нуля.

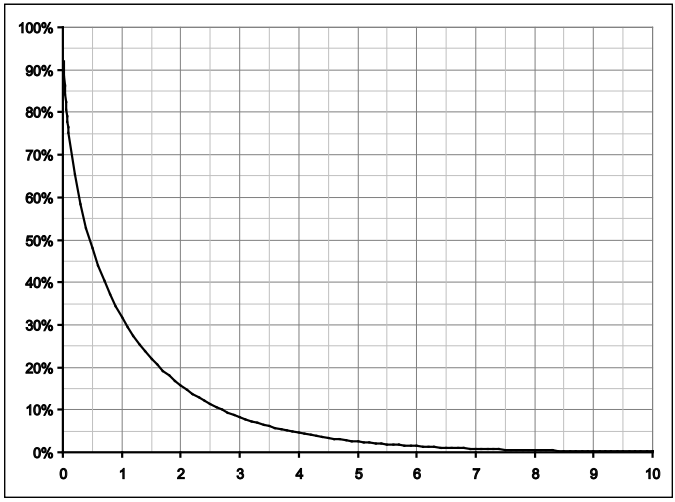


Рис. 1. Значимость связи для данных табл. 1

5%-ный порог значимости соответствует значению статистики, равной 3,8415, а для наших данных ($\chi^2 = 3{,}8429$) значимость равна 4,996%.

Для оценки масштаба отклонений наблюдаемых частот от ожидаемых в каждой клетке таблицы проводят стандартизацию остатков [2], деля отклонение на стандартную ошибку.

$$z_{ij} = \frac{(n_{ij} - e_{ij})}{s}. \qquad (6)$$

Случайная величина z_{ij} с достаточно хорошим приближением подчиняется стандартному нормальному закону распределения. Стандартная ошибка s приближенно рассчитывается через оценку дисперсии:

$$s^2 = \frac{n_{i\circ} \cdot (N - n_{i\circ}) \cdot n_{\circ j} \cdot (N - n_{\circ j})}{N \cdot N \cdot (N - 1)}. \quad (7)$$

Если $s^2 < 9$, то пользуются точной оценкой дисперсии, вычисляемой из гипергеометрического распределения n_{ij}. Если критерий χ^2 выявляет наличие связи по таблице сопряженности в целом, то стандартизованные отклонения позволяют уточнить, какие именно клетки в этой таблице вносят наибольший вклад в обнаруженную связь.

Как видно из табл. 1, в нашем случае стандартизованное отклонение (1,96) тоже примерно соответствует 5%-ному уровню значимости, как и статистика хи-квадрат. Это практически полное совпадение значимости, оцененной разными способами, укрепляет доверие к получаемым результатам.

Для определения значимости связи по таблице сопряженности минимального размера (2×2), кроме критерия хи-квадрат, можно применять так называемое *гипергеометрическое распределение.* Оно позволяет вычислить вероятность наблюдения случайной величины ξ в условиях нулевой гипотезы:

$$p_{n_{11}} = P(\xi = n_{11}) = \frac{C_{n_{1\circ}}^{n_{11}} C_{n_{2\circ}}^{n_{21}}}{C_N^{n_{\circ 1}}}. \quad (8)$$

Гипергеометрическое распределение, в отличие от *хи-квадрат*, является точным. Чтобы понять, как получается формула (8), построим наглядную модель. Пусть имеется совокупность, содержащая N объектов двух типов. Число объектов первого типа – $n_{1\circ}$, а второго – $n_{2\circ} = N - n_{1\circ}$. Далее ту же совокупность делим на две группы. Первая из них содержит $n_{\circ 1}$ объектов, а вторая – $n_{\circ 2} = N - n_{\circ 1}$.

Теперь представим модель в виде таблицы 2×2. В клетку (1,1) заносим число n_{11} объектов первого типа, попавших в пер-

вую группу, т.е. тех, чьи порядковые номера в совокупности оказались не больше $n_{\circ 1}$. В клетку (1,2) заносим число n_{12} объектов первого же типа, но попавших во вторую группу, т.е. тех, чьи порядковые номера больше $n_{\circ 1}$. Соответственно интерпретируются частоты n_{21} и n_{22}.

Так как любая из частот $\{n_{11}, n_{12}, n_{21}, n_{22}\}$ однозначно определяет три остальные, таблице соответствует одна степень свободы. Не теряя общности, предположим, что независимой величиной будет n_{11}. Чтобы построить гипергеометрическое распределение, нам нужно найти вероятность сложного события – попадания в точности n_{11} объектов первого типа в группу объема $n_{\circ 1}$ из совокупности объема N. Сложным событие называется потому, что оно складывается из множества элементарных. В данном случае элементарным событием является одна из возможных уникальных комбинаций из N объектов, расположенных на N упорядоченных местах. Полное количество всех элементарных событий равно хорошо известному в комбинаторике числу *перестановок* $P_N = N!$. Примем классическое для теории вероятностей предположение, что все эти элементарные события равновероятны, поскольку нет оснований считать некоторые из перестановок более вероятными, чем другие.

Среди всего множества элементарных событий будем рассматривать только те, которые содержат в точности n_{11} объектов первого типа на $n_{\circ 1}$ первых местах и, соответственно, n_{12} объектов на $n_{\circ 2} = N - n_{\circ 1}$ последних местах. Подсчитаем точное количество таких событий.

Для этого найдем число уникальных *сочетаний* (без учета порядка расположения) n_{11} объектов первого типа в первой группе: $C_{n_{1\circ}}^{n_{11}} = \dfrac{n_{1\circ}!}{n_{11}!n_{12}!}$. Аналогично для объектов второго типа в той же группе: $C_{n_{2\circ}}^{n_{21}} = \dfrac{n_{2\circ}!}{n_{11}!n_{12}!}$.

Также нам понадобится полное число *перестановок* всех $n_{\circ 1}$ объектов внутри первой группы: $P_{n_{\circ 1}} = n_{\circ 1}!$ и оставшихся $n_{\circ 2} = N - n_{\circ 1}$ объектов внутри второй группы: $P_{n_{\circ 2}} = n_{\circ 2}!$

Окончательно число элементарных событий равно произведению всех упомянутых выше *сочетаний* и *перестановок*: $\frac{n_{1\circ}!n_{2\circ}!n_{\circ 1}!n_{\circ 2}!}{n_{11}!n_{12}!n_{21}!n_{22}!}$. Отсюда сразу получаем вероятность события как отношение числа:

$$P_{2\times 2}(\xi = n_{11}) = \frac{n_{1\circ}!n_{2\circ}!n_{\circ 1}!n_{\circ 2}!}{n_{11}!n_{12}!n_{21}!n_{22}!N!}. \qquad (9)$$

Нетрудно убедиться, что формулы (8) и (9) идентичны, но при этом формулу (9) легко распространить на случай таблицы произвольного размера $r \times c$:

$$P_{r\times c}(\xi = \{n_{ij}\}) = \frac{\prod_{i=1}^{r} n_{i\circ}! \prod_{j=1}^{c} n_{\circ j}!}{\prod_{i=1}^{r}\prod_{j=1}^{c} n_{ij}!N!}. \qquad (10)$$

Проблемой при использовании такого «обобщенного» гипергеометрического распределения является то, что число степеней свободы $\nu = (r - 1)(c - 1)$ для произвольной таблицы может быть больше единицы и потому для построения функции распределения необходим специальный достаточно сложный алгоритм, осуществляющий последовательный перебор и упорядочение по вероятностям всех возможных вариантов заполнения таблицы. В работе [3] описан *полнопереборный* вариант такого алгоритма. Для больших таблиц сопряженности он крайне неэффективен по времени счета. Возможно, в дальнейшем удастся оптимизировать схему вычисления до такой степени, что точный расчет значимости по «обобщенному» гипергеометрическому распределению сможет заменить приближенный по критерию *хи-квадрат*. По крайней мере, заменить в тех случаях, когда критерий *хи-квадрат* оказывается неприменим. На наших данных доля таких случаев составила более 40%.

Связь между переменными, имеющими номинальный и интервальный уровни измерения

Для случая такой связи можно воспользоваться дисперсионным анализом [1, с. 166]. Статистика, на основе которой проверяется гипотеза о наличии связи, вычисляется как отношение межгрупповой дисперсии к внутригрупповой.

$$F = \frac{\frac{1}{k-1}\sum_{j=1}^{k} n_j (x_{\bullet j} - \bar{x})^2}{\frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_j - x_{\bullet j})^2}, \qquad (11)$$

где N – объем выборки (число объектов); k – число групп (различных значений номинальной переменной); n_j – число объектов в группе j; $x_{\bullet j}$ – среднее значение интервальной переменной по этой группе; $\bar{x}$ – среднее значение по выборке.

Эта статистика может быть проинтерпретирована как отношение сигнала к шуму. Нулевая гипотеза формулируется как равенство нулю этой статистики, что эквивалентно равенству всех средних по группам между собой и среднему по выборке, т.е. независимости интервальной переменной от номинальной. Значимость связи рассчитывается по распределению Фишера с $(k-1, N-k)$ степенями свободы.

На рис. 2 приведен результат проверки связи оценки респондентом соотношения модернистов-рационалистов в современном казахстанском обществе с положением респондента на работе. Видно, что самый оптимистичный взгляд на количество сторонников модернизации проявляют директора и заместители директоров, а самый пессимистичный – служащие среднего звена. На графике для каждого ответа переменной X представлено среднее значение в группе по переменной Y и стандартная ошибка среднего. Число показывает объем группы – количество респондентов, выбравших этот ответ. Чем больше группа, тем меньше ошибка среднего. Пунктиром в виде коридора вокруг среднего показано стандартное отклонение среднего по выборке.

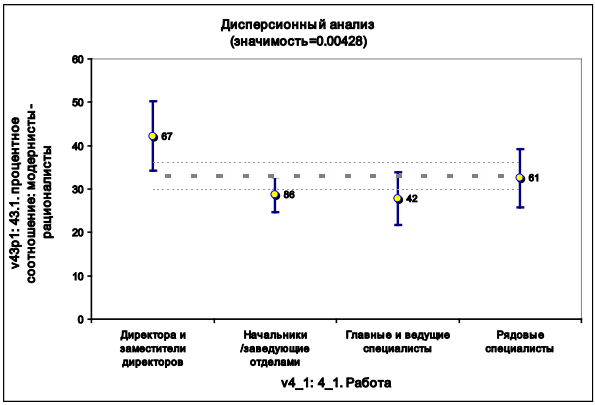


Рис. 2. Результаты дисперсионного анализа

Общая дисперсия выборки в данном случае составляет 167 901,9, которая распадается на межгрупповую 8 468,8 и внутригрупповую 159 433,1. Кажется, что межгрупповая дисперсия намного меньше внутригрупповой, но при вычислении статистики учитывается и число степеней свободы для каждого вида дисперсии, которых для межгрупповой дисперсии только 3 (число групп минус единица), а для внутригрупповой – 254. С учетом этого межгрупповая дисперсия, приходящаяся на одну степень свободы, равна 2 822,9, а внутригрупповая (остаточная) равна 627,6. Отношение дисперсий равно 4,5, что дает значимость связи – 0,004. Из этого можно сделать вывод, что между оценкой экспертом соотношения модернистов и рационалистов в казахстанском обществе и служебным положением самого эксперта существует связь.

Связь переменных, имеющих интервальный уровень измерения

Как известно, коэффициент корреляции Пирсона [1, с. 263–265] характеризует простейший вид зависимости между двумя количественными переменными – линейную связь:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \qquad (12)$$

где x, y – количественные переменные; N – объем выборки.

Коэффициент корреляции R показывает тесноту линейной связи двух переменных, но по его величине нельзя ничего сказать о статистической значимости связи. Для этого в статистике используют случайную величину T, которая подчиняется распределению Стьюдента с $(N - 2)$ степенями свободы.

$$T = \pm\sqrt{\frac{(N-2)R^2}{1-R^2}}, \text{ здесь знак } T \text{ совпадает со знаком } R. \qquad (13)$$

Поскольку в социологических исследованиях переменные количественного характера реже встречаются, чем порядковые, то вместо коэффициента корреляции Пирсона применяется коэффициент корреляции Спирмена, который отличается тем, что вместо самих значений x, y используются их ранги.

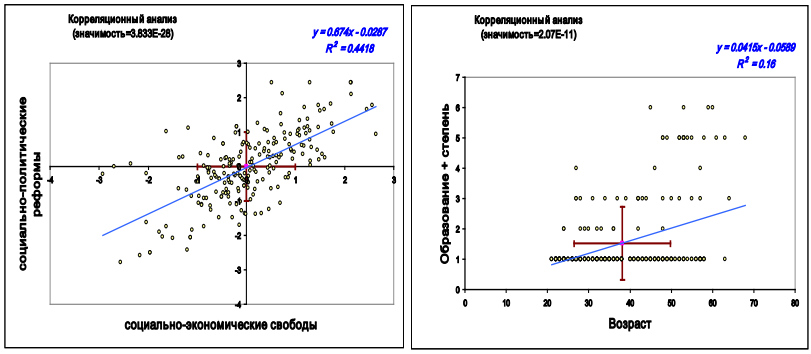


Рис. 3. Результаты корреляционного анализа для двух факторов (вверху) и для количественной и балльной переменной (внизу)

Оценки значимости связи статистическими экспериментами

Из определения значимости как вероятности получить вычисленную или большую величину статистики при выполнении нулевой гипотезы следует, что сама значимость в условиях нулевой гипотезы должна быть распределена равномерно от нуля до единицы. Это позволяет проверить применимость любого метода анализа связи для любой пары переменных. Для этого необходимо построить эмпирическое распределение значимости связи для пары переменных при выполнении условий нулевой гипотезы.

В нашем случае условия нулевой гипотезы реализуются простейшим способом – перемешиванием наблюдений в одной из двух переменных, т.е. изменением порядка следования анкет в одном из столбцов матрицы данных с помощью генератора случайных чисел. При этом полностью сохраняются одномерные распределения обеих переменных, но связь между ними уничтожается. Каждый такой эксперимент дает одну реализацию случайной величины – *значимости связи*. Проведя достаточно много статистических экспериментов, мы получаем эмпирическое распределение этой величины в условиях нулевой гипотезы. Если оно окажется близким к равномерному, то будет основание утверждать, что проверяемая статистика для выбранной пары переменных на имеющейся выборке *работает* корректно.

На рис. 4 показаны случаи, когда проверка дает положительный результат. При этом эмпирическое распределение выглядит одинаково для всех методов. На рис. 5 приведены случаи с неудовлетворительным исходом, и каждый метод проявляется по-своему.

На нашем массиве чаще всего неприменимыми оказывались статистика хи-квадрат (20 122 случая из 49 141, что составляет 40,9%), реже – дисперсионный анализ (2 114 случаев из 14 892, что соответствует 14,2%) и реже всего – корреляция (71 случай из 1 149, что равно 6,2%). В случае с корреляцией наблюдаемый процент отрицательных проверок очень близок к пороговому зна-

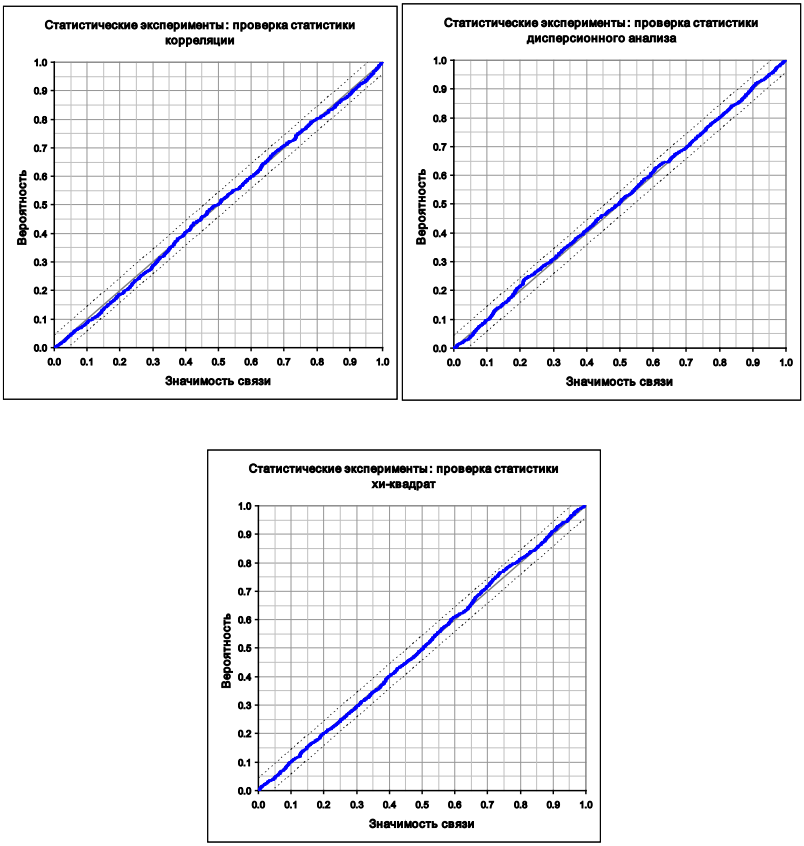


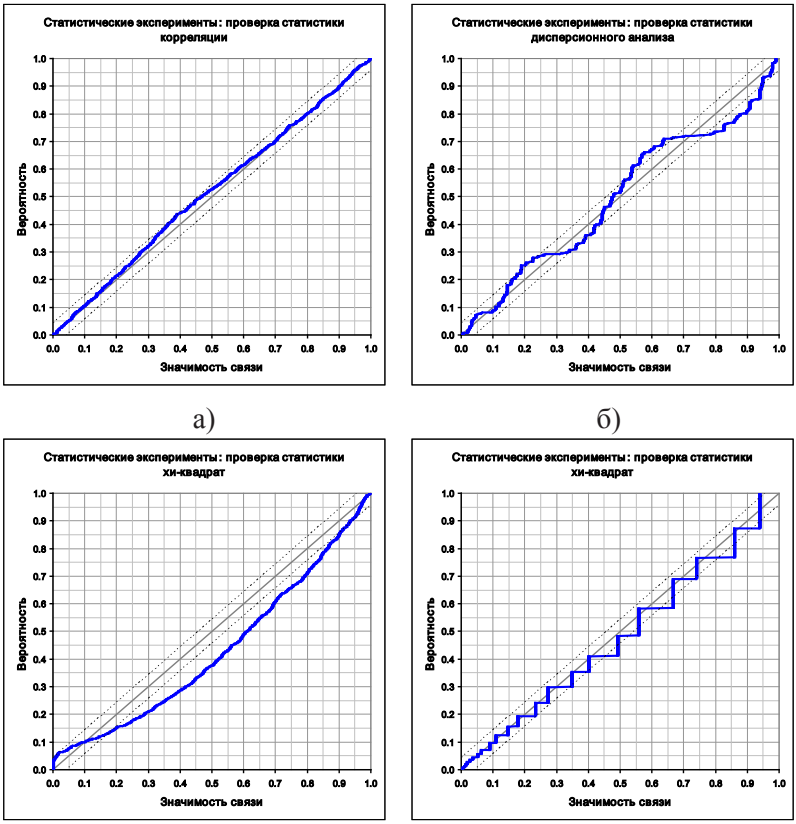
Рис. 4. Сравнение распределения значимости связи с равномерным распределением. Штриховыми линиями показан 95%-ный доверительный интервал (по критерию Колмогорова-Смирнова)

чению для отсева – 5%. Поэтому оснований говорить о неприменимости какой-либо из гипотез нет. На рис. 5а показано типичное для корреляции распределение, которое нехарактерно для систематических отклонений. Тем более, что повторная проверка всех случаев не выявила ни одной подозрительной гипотезы.

В случае дисперсионного анализа наблюдалась более характерная и устойчиво повторяющаяся картина, такая как на рис. 5б. Здесь одна из переменных является дихотомической, причем единичное значение встречается только в 4 анкетах из 260.

На рис. 5в видно, что статистика хи-квадрат для таблицы сопряженности чаще ожидаемого принимает значения, близкие к нулю и к единице, и реже – промежуточные. Это вызвано тем, что в таблице сопряженности (см. табл. 2) содержится много клеток, ожидаемые частоты в которых близки к нулю. В результате этого наиболее вероятные нулевые значения в этих клетках порождают заниженные значения статистики, а ненулевые – завышенные.

Известно эмпирическое правило для проверки применимости метода хи-квадрат [4], которое требует, чтобы ожидаемые частоты во всех ячейках таблицы сопряженности были не менее 1 и в 80% клеток – не менее 5. Из табл. 2 видно, что это условие грубо нарушается. Однако справедливости ради надо заметить, что далеко не всегда нарушение этого правила сопровождается отклонением распределения значимости от равномерного. Таким образом, предлагаемая нами проверка применимости метода часто оказывается более мягкой, т.е. позволяет проверять связи по таблицам с малыми ожидаемыми частотами.



а) б)

в) г)

Рис. 5. Примеры отрицательного исхода проверки корректности

Таблица 2

ПРИМЕР ТАБЛИЦЫ СОПРЯЖЕННОСТИ С МАЛЫМИ ОЖИДАЕМЫМИ ЧАСТОТАМИ

v11: На Ваш взгляд, к какому типу относится современное казахстанское общество?	v22: Оцените ход модернизации казахстанского общества				
	Положительно	Скорее положительно	Скорее отрицательно	Отрицательно	Итого
Современное модерное	19,88	39,75	7,06	1,31	68
Скорее современное	41,22	82,43	14,64	2,71	141
Скорее традиционное	9,35	18,71	3,32	0,62	32
Традиционное	1,46	2,92	0,52	0,10	5
Смешанное	2,92	5,85	1,04	0,19	10
Переходное	0,58	1,17	0,21	0,04	2
В равной мере и современное, и традиционное	0,29	0,58	0,10	0,02	1
На пути европеизации с элементами национальных культур	0,29	0,58	0,10	0,02	1
Итого	76	152	27	5	260

Таблица 3

ФРАГМЕНТ ТЕОРЕТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ ЗНАЧИМОСТИ

n_{11}	χ^2		Показатели гипергеометрического распределения		
	Статистика	Значимость, %	Вероятность, %	Значимость, %	Значимость*, %
38	3,84	5,00	1,50	5,66	4,91
39	2,91	8,81	2,39	9,86	8,66

Окончание табл. 3

n_{11}	χ^2		Показатели гипергеометрического распределения		
	Статистика	Значимость, %	Вероятность, %	Значимость, %	Значимость*, %
40	2,11	14,68	3,57	16,24	14,45
41	1,43	23,16	4,99	25,31	22,82
42	0,89	34,65	6,54	37,42	34,15
43	0,47	49,22	8,03	52,57	48,55
44	0,19	66,57	9,25	70,34	65,72
45	0,03	85,93	9,98	89,90	84,91
46	0,01	93,83	10,10	100,00	100,00
47	0,11	73,98	9,58	79,92	75,13
48	0,34	55,73	8,52	61,10	56,83
49	0,71	40,00	7,11	44,54	40,98
50	1,20	27,30	5,57	30,88	28,10
51	1,82	17,67	4,09	20,32	18,28
52	2,58	10,83	2,81	12,67	11,26
53	3,46	6,28	1,82	7,47	6,56

* Скорректированное (квазинепрерывное) гипергеометрическое распределение.

Особый случай, когда проверка по эмпирическому распределению значимости может приводить к ошибочному выводу, показан на рис. 5г. Таблица сопряженности для этого случая (табл. 2) не содержит клеток с малыми ожидаемыми частотами, но она выделяется тем, что в ней всего четыре клетки с одной степенью свободы. Для нее нетрудно построить точное теоретическое распределение значимости (в табл. 3 приведен его фрагмент).

На рис. 6 приведены теоретическое и эмпирическое распределения, полученные по результатам 1000 статистических экспериментов. Они практически идентичны, а теоретическое не вписывается в границы 5%-ного доверительного интервала, что является следствием дискретного характера распределения, вызванного малым числом степеней свободы.

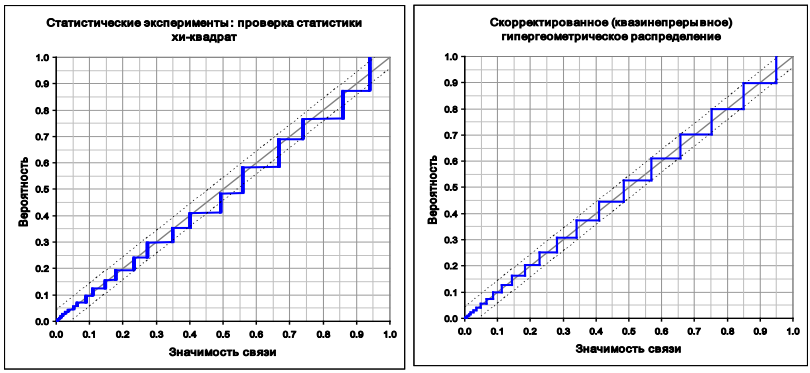


Рис. 6. Эмпирическое и теоретическое распределения

Поскольку теоретическое распределение уже построено, можно сравнить его с распределением хи-квадрат. На рис. 7 представлены результаты сравнения вероятности наблюдения частоты n_{11}, рассчитанной методом χ^2 (черные точки) и с помощью гипергеометрического распределения (белые ромбики). На рис. 7б видно, что квазинепрерывная корректировка гипергеометрического распределения делает его существенно ближе к распределению χ^2.

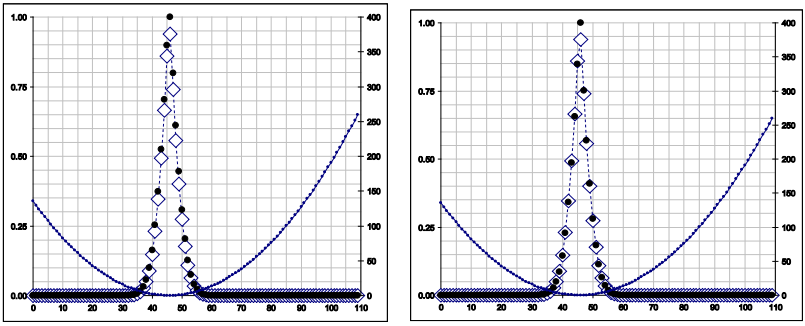


Рис. 7. Сравнение распределения хи-квадрат с гипергеометрическим

Параболой на графике отображено значение статистики *хи-квадрат*, которая при $n_{11} = 109$ равна 260. Жирными точками показана значимость нулевой гипотезы при расчете через гипергеометрическое распределение, штриховой линией с ромбами – значимость по хи-квадрат. Видно, что распределение хи-квадрат близко к гипергеометрическому, а наблюдающиеся отличия можно объяснить тем, что гипергеометрическое распределение абсолютно точное дискретное, а хи-квадрат – непрерывное и приближенное. Если сравнить значимости для $n_{11} = 38$, то можно убедиться, что они в обоих случаях достаточно близки: для хи-квадрат – 5,00%, для гипергеометрического распределения – 5,66% (случай дискретного) и 4,91% (случай скорректированного квазинепрерывного).

Чтобы получить значимость нулевой гипотезы для заданного n_{11} по гипергеометрическому распределению, необходимо из 100% вычесть вероятность всех более «вероятных» значений n_{11}. Например, $P(n_{11} = 47) = 100\% - 10{,}1\% - 9{,}98\% = 79{,}92\%$. Корректировка значимости сводится к тому, что мы дополнительно вычитаем половину вероятности самого значения n_{11}:

$$\mathrm{P}(n_{11} = 47) = 100\% - 10{,}1\% - 9{,}98\% - 9{,}58\% : 2 = 75{,}13\%.$$

Эта половина берется из тех соображений, что мы превращаем дискретное распределение в непрерывное, в результате чего точное значение n_{11} равномерно рассеивается в окрестности 47. При этом половина значений становится больше 47, а половина – меньше. В случае наиболее вероятного значения (46) корректировка не требуется, так как в экстремуме первая производная плотности распределения равна нулю и отклонение в любую сторону от 46 не увеличивает плотность вероятности. Очевидно, что предложенная корректировка является достаточно грубой. Ее уточнение требует аккуратной аппроксимации дискретного распределения непрерывным.

Следует особо подчеркнуть, что сплошной, без пропусков, просмотр упорядоченного по значимости *списка парных связей* дает исследователю возможность подойти к анализу данных непредвзято, обнаружить не только то, что он готов увидеть в соответствии со своими теоретическими представлениями и априорными гипотезами, но и связи, которые действительно существуют в эмпирическом материале. Необходимость объяснения подобных связей наталкивает исследователя на формулирование новых содержательных гипотез. Например, при рассмотрении результатов дисперсионного анализа могут появиться идеи построения количественных переменных на базе некоторых номинальных методами оцифровки данных [5, с. 344].

В процессе сплошного поиска важным является формирование множества проверяемых гипотез о связях. Выше отмечалось, что необходимо отклонять те гипотезы, для которых статистика связи не работает. Но, кроме того, должны быть отклонены еще и те гипотезы, которые дают тривиальный или содержательно предсказуемый результат. Например, если вычислять корреляцию между взаимно ортогональными по построению факторами, то можно априори сказать, что она будет в точности нулевой. Также не имеет смысла искать связь между возрастом исходным и возрастом, укрупненным по интервалам в 5 или 10 лет. Результат не будет представлять никакого интереса.

Проверка гипотезы об отсутствии связей в массиве данных

На рис. 8 представлены результаты отсева некорректных гипотез. Нижние кривые – до отклонения некорректных гипотез, верхние – после. Видно, что в нашем примере распределение значимости парных связей в массиве данных после отсева смещается в сторону увеличения количества статистически значимых связей. После того, как мы оставили только корректные гипотезы, появляется возможность проверить гипотезу о наличии парных связей в массиве данных в целом. Для этого достаточно воспользоваться проверкой совпадения распределения значимости с равномерным по критерию Колмогорова-Смирнова.

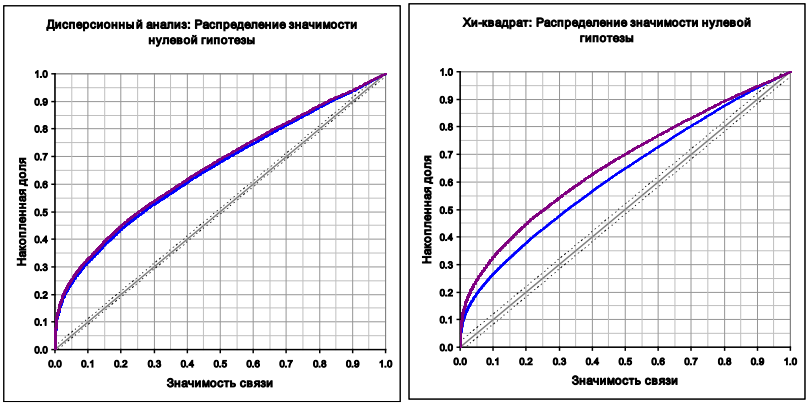


Рис. 8. Распределение значимости нулевой гипотезы

По итогам проведенных экспериментов возможны следующие выводы:

• Увеличение мощности компьютеров позволяет ставить и решать задачи, которые раньше казались невыполнимыми из-за большого объема вычислений.

• Одним из классов таких задач является проверка сложных статистических гипотез на исходных для анализа данных посредством проведения вычислительных экспериментов.

• Предложенная методика позволяет существенно повысить адекватность статистического анализа данных, подвергая проверке не только содержательные предположения о наличии парных связей между переменными, но и оценить работоспособность самого статистического критерия, с помощью которого проводится анализ.

• Использование предложенных методических приемов позволит существенно поднять качество опросов, создавая возможность проведения статистической экспертизы результатов социологических опросов.

ЛИТЕРАТУРА

1. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. М.: Большая российская энциклопедия, 1999.

2. *Костин В.С.* Статистика для сравнения классификаций // Информационные технологии в гуманитарных исследованиях: Сб. тр. Новосибирск, 2003. С. 57–65. Вып. 6.

3. http://www.sati.archaeology.nsc.ru/sibirica/Data/int6/?html=int67.htm&mi=izdaniya&id=1826.

4. *Haberman Sh.J.* Analysis of Qualitative Data. N.Y.: Academic Press, 1978. Vol. 1.

5. SPSS Base 8.0 для Windows. М.: Изд-во Центра общечеловеческих ценностей, 1998.