УДК 519.22 / .25: 94 (100) “1939/1945”

В.Ю. Дубницкий, А.И. Ходырев

Харьковский институт банковского дела Университета банковского дела Национального Банка Украины (г. Киев)

СТАТИСТИЧЕСКИЙ АНАЛИЗ ПРИЧИН АНТИСОВЕТСКИХ НАСТРОЕНИЙ У ПЕРЕМЕЩЁННЫХ ЛИЦ ПО ДАННЫМ ГАРВАРДСКОГО ПРОЕКТА

Для анализа многоклеточных таблиц (категоризированных данных) предложена двухступенчатая процедура, отличающаяся тем, что на первой ступени использована диагностика Симонова-Цая для оценки возможности последующего анализа, основанного на использовании величины X^2. Сравнение полученных результатов проверки статистической гипотезы об отсутствии зависимости между фактором строк и фактором столбцов в многоклеточной таблице показало, что критерии её проверки X^2, отношения правдоподобия, Кресси-Рида, Хеллингера и Зелтермана дают близкие результаты. Для вычислений использованы данные анкетирования перемещенных лиц, полученные в результате реализации гарвардского проекта в 1948-1951 г.г. Установлено, что степень неприятия советской власти перемещёнными лицами не зависела от их национального состава, уровня квалификации, принадлежности к одной из социальных групп, особенностей личной биографии в советский период.

Ключевые слова: *категоризированные данные, многоклеточные таблицы, проверка статистических гипотез, диагностика Симонова-Цоя, критерий хи-квадрат, критерий отношения правдоподобия, критерий Кресси-Рида, критерий Хеллингера, критерий Зелтермана.*

Введение

Анализ таблиц сопряжённости признаков нашёл широкое применение в социологии, психологии, медицине и других областях знаний, в которых результаты исследований представлены в нечисловой шкале признаков (категоризированные данные) [1, 2]. Анализ многоклеточных таблиц получил широкое распространение потому, что даёт возможность проверить методами статистической теории гипотез содержательные нестатистические гипотезы, относящиеся к различным предметным областям. За последние годы появились исследования, в которых описаны новые методы, предназначенные для такого анализа. В настоящей работе описано применение этих методов на примере анализа данных, полученных в результате изучения результатов Гарвардского проекта.

Анализ источников исходных статистических данных

Термин «перемещённые лица» или «displaced persons» возник после окончания Второй мировой войны и применялся по отношению к лицам, насильственно вывезенным в Германию. Большинство советских граждан, подпадавших под эту категорию, были добровольно или принудительно репатриированы в СССР. Меньшая часть (около 250 тыс. человек [3]) отказалась вернуться в СССР. За ними был закреплён термин «DP-persons».

В марте 1948 года в США было решено для заполнения ниши в разведывательной деятельности по добыванию сведений о реальном устройстве жизни в СССР и системе принятия решений советскими руководителями любого уровня провести опрос среди перемещённых лиц. Цель опроса – заполнение пробелов в сведениях по указанной тематике. Всего было опрошено 12,5 тыс. человек. После публикации материалов на сайте http://oasis.lib.harvard.edu/oasis/deliver/deepLink?_collection=oasis&uniqueId=fun00001 они вошли в широкий научный оборот под названием «Гарвардский проект».

В 1941 г. при вступлении в Смоленск немецкими войсками был захвачен архив Смоленского областного комитета ВКП(б). После окончания Великой Отечественной войны он также оказался в распоряжении властей США и в 2002 г. был возвращён в Россию. В исторической литературе этот материал известен как «Смоленский архив». Сегодня именно эти два собрания документов считают наиболее важными источниками данных при исторических и социологических исследованиях различных сторон жизни СССР в предвоенные и военные годы. В 2003 г. вышла работа [4], в которой приведены структурированные результаты интервью, включённых в материалы проекта. Следует отметить, что автор работы [4] настолько тщательно отнёсся к систематизации данных, что кроме своих вариантов группировки привёл, с подробным описанием, альтернативные варианты. Организаторами и основными исполнителями проекта были американские военные специалисты, «социологи в штатском» и сотрудники корпорации RAND. Естественно, что реализация проекта была бы невозможна без участия советских граждан из категории DP. Сведения о них приведены в [4,

с. 57…59]. Общее в их биографиях то, что все они занимали высокое, по меркам тех лет, общественное положение и не принадлежали к слою управленцев. Практически неструктурированный исходный материал и сложности его обработки привели к тому, что между разными разделами цитируемой работы имеются противоречия. Исходный материал для статистических исследований приведён в табл. 2 – 5.

Таблица 1

Особенности биографии советских граждан-участников Гарвардского проекта

Фамилии, должность, занимаемая в СССР до начала Великой отечественной войны	Особенности биографии			
	Арестован НКВД	Сотрудничал с оккупантами	Работал на американскую военную разведку (DIA)	Сотрудничал с ЦРУ (CIC)
Яковлев Б.А., вице-президент Академии архитектуры	+	+	*	*
Кунта А.А., (Авторхан Авторханов**))	+	+	+	
Алдан М.А., полковник Красной Армии	*	+	+	+
Штеппа К.Ф., профессор-историк	+	+	+	+
Криптон К.Г., начальник отдела НИИ	*	+	+	+
Филиппов А.П., профессор-философ	*	+	+	+
Марченко В.П., научный сотрудник	*	*	+	+
Ниеман Ю.М., научный сотрудник	+	*	+	+

Примечание: * – сведения отсутствуют;
** – современному читателю А. Авторханов известен своей работой [9].

Таблица 2

Степень неприятия советской системы разными слоями населения (%)

Степень неприятия	Интеллигенция	Служащие	Высококвалифицированные рабочие	Простые рабочие	Колхозники
Низкая	71	59	54	42	40
Средняя	17	25	26	24	24
Высокая	12	16	20	34	36
Всего (чел.)	567	607	243	410	312

Исходные данные: [4, с. 116].

Таблица 3

Процент социальных групп, выражавших разную степень враждебности по отношению к режиму

Социальная группа	Процент заявивших, что:			Количество респондентов (чел.)
	Ленин принёс много вреда	Большевистские лидеры достойны смерти	Большевики хуже	
Интеллигенция	74	39	70	622
Служащие	78	46	68	665
Высококвалифицированные рабочие	70	47	60	268
Простые рабочие	73	54	51	458
Колхозники	72	60	44	348

Исходные данные: [4, с. 118].

Таблица 4

Процент представителей разных социальных групп, выражавших неприятие по причине арестов

Опыт ареста	Интеллигенция	Служащие	Высококвалифици-рованные рабочие	Простые рабочие	Колхозники
Арестован сам или член семьи	35	43	51	68	67
Не было арестов	12	24	33	30	49

Исходные данные: [4, с. 123].

Таблица 5

Неприятие советской системы по национальному признаку среди разных социальных групп населения (%)

Национальность	Интеллигенция	Служащие	Высококвалифициро-ванные рабочие	Простые рабочие	Колхозники
Русские	28	42	48	57	67
Украинцы	29	43	47	63	60
Другие	27	33	36	47	55

Источник: [4, с. 125].

В работе [4, с. 120] приведена весьма подробная таблица, в которой даны сведения о причинах неприятия советского режима разными социальными группами населения. Не все приведённые в работе [4] таблицы дают возможность их корректного анализа. В некоторых таблицах, например, в табл. 2, есть возможность прямого перехода от относительных данных, выраженных в процентах к абсолютным данным, в остальных случаях авторы предлагаемого сообщения проводили специальный расчёт. Разумеется, что данное обстоятельство ни в коем случае не может быть поставлено в упрёк автору исследования [4], который проделал огромную работу по вводу этих данных в научный оборот.

На основе данных, приведённых в [4, с. 125], составлена табл. 6, содержащая информацию о социально-классовом составе опрошенных DP из числа советских граждан.

Таблица 6

Группировка участников анкетирования по признаку социальной принадлежности

Социально-классовая группа	Интеллигенция	Служащие	Высококвалифицированные рабочие	Простые рабочие	Колхозники
Всего человек	721	698	295	519	385

Сведения о количественном составе социальных групп, принявших участие в опросе на тему оценки степени неприятия советской системы, приведены в табл. 7. Количественный состав социальных групп, выражавших разную степень враждебности по отношению к режиму, показан в табл. 8.

Таблица 7

Степень неприятия советской системы разными слоями населения

Степень неприятия	Интеллигенция	Служащие	Высококвалифицированные рабочие	Простые рабочие	Колхозники
Низкая	403	358	131	172	125
Средняя	96	152	63	99	75
Высокая	87	97	49	139	112

Составлено авторами по данным табл. 2.

Таблица 8

Количественный состав социальных групп, выражавших разную степень враждебности по отношению к режиму

Социальная группа	Процент заявивших, что:		
	Ленин принёс много вреда	Большевистские лидеры достойны смерти	Большевики хуже
Интеллигенция	252	133	238
Служащие	270	159	236
Высококвалифицированные рабочие	106	71	91
Простые рабочие	188	139	131
Колхозники	142	119	87

Рассчитано авторами по данным табл. 3.

Таблица 9

Количественный состав социальных групп, выражавших своё неприятие по причине арестов

Опыт ареста	Интеллигенция	Служащие	Высококвалифицированные рабочие	Простые рабочие	Колхозники
Арестован сам или член семьи	249	300	150	363	258
Не было арестов	30	72	50	106	126

Рассчитано авторами по данным табл. 4.

Таблица 10

Количественные данные о национальном составе лиц, заявивших о неприятии советской системы

Национальность	Интеллигенция	Служащие	Высококвалифицированные рабочие	Простые рабочие	Колхозники
Русские	240	248	108	177	142
Украинцы	249	254	106	196	127
Другие	232	195	81	146	385

Рассчитано авторами по данным табл. 5.

Анализ литературы по статистическому анализу категоризированных данных

Основные, классические методы обработки данных, представленных в категоризированном виде, приведены в работах [2 – 4]. Исходные данные представляют в виде таблицы размерности $r \times c$, то есть таблицы, содержащей r строк и c столбцов. Общий вид такой таблицы, получившей название таблицы сопряжённости, показан в табл. 11.

При составлении этой таблицы приняты следующие условные обозначения. A_{ij} – количество объектов, которым присущи признаки i и j одновременно. Эту величину называют фактической частотой таблицы сопряжённости.

Таблица 11

Представление категоризированных данных в виде $r \times c$ таблицы

Признак строки	Признак столбца						
	1	2	…	j	…	c	$\sum$
1	A_{11}	A_{12}	…	A_{1j}	…	A_{1c}	$n_{1\bullet}$
2	A_{21}	A_{22}	…	A_{2j}	…	A_{2c}	$n_{2\bullet}$
…	…	…	…	…	…	…	…
i	A_{i1}	A_{i2}	…	A_{ij}	…	A_{ic}	$n_{i\bullet}$
…	…	…	…	…	…	…	…
r	A_{r1}	A_{r2}	…	A_{rj}	…	A_{rc}	$n_{r\bullet}$
$\sum$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet c}$	$n_{\bullet\bullet} = n$

Суммы по строкам таблицы сопряжённости определяют по формуле:

$$n_{i\bullet} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...r . \quad (1)$$

Суммы по столбцам таблицы сопряжённости определяют по формуле:

$$n_{\bullet j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...c . \quad (2)$$

Общее количество наблюдаемых объектов определяют по формуле:

$$n = \sum_{k=1}^{r}\sum_{l=1}^{c} A_{kl} . \quad (3)$$

Основная статистическая гипотеза, проверяемая при анализе таблиц сопряжённости – гипотеза об отсутствии взаимодействия между фактором, задаваемым уровнями строк и фактором, задаваемым уровнем столбцов. Если это взаимодействие отсутствует, то исходные данные должны быть распределены равномерно. Для проверки этой гипотезы чаще всего вычисляют величину хи-квадрат по формуле:

$$X^2 = \left(A_{ij} - E_{ij}\right)^2 \Big/ E_{ij} , \quad (4)$$

где E_{ij} – ожидаемое количество (частота) объектов, имеющих признаки i и j одновременно в предположении справедливости равномерного распределения:

$$E_{ij} = n_{i\bullet} n_{\bullet j} / n . \quad (5)$$

Если это предположение, принимаемое в качестве гипотезы H_0, справедливо, то величина хи квадрат имеет распределение χ^2 с $\nu = (r-1)(c-1)$ степенями свободы. Решающее правило S принятия гипотез имеет вид:

$$S = \begin{cases} H_0, \text{если} \quad \chi^2 < \chi_{\alpha,f}; \\ H_1, \text{если} \quad \chi^2 > \chi^2_{\alpha,f}, \end{cases} \quad (6)$$

где α – принятый уровень доверительной вероятности.

На этом, как правило, заканчивается анализ таблиц сопряжённости в той последовательности, которая изложена в работах [1 – 3, 5 – 7] и известного как критерий Хи-квадрат. В дальнейшем изложении этот критерий будем обозначать символом критерий К2.

В работе [8] приведены результаты по анализу таблиц сопряжённости, полученные в последние годы и реализованные в программной системе AtteStat. Начинать анализ таблиц сопряжённости автор работы [8] рекомендует начинать с использования диагностики Симонова-Цая (критерий К1). Для этого вычисляют величину:

$$S = \frac{\left(\chi^2(\nu,\alpha)\right)^{1/2}}{3\left(X^2\right)^{3/2}} \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left|\left(A_{ij} - E_{ij}\right)\right|^3}{E_{ij}^2} . \quad (7)$$

В условии (7) принято, что $\chi^2(\nu,\alpha)$ – значение обратной функции распределения χ^2 с ν степенями свободы и уровнем доверительной вероятности α. Если величина S > 0,25, то использование критериев, основанных на применении величины X^2 не рекомендуется.

Критерий отношения правдоподобия (К3) имеет вид:

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} A_{ij} \ln \frac{A_{ij}}{E_{ij}} . \quad (8)$$

При применении этого критерия принято, что если A_{ij} =0, то есть наблюдения с данным сочетанием факторов отсутствуют, то значение критерия для этой клетки таблицы принимают равным нулю. Величина G^2 распределена в соответствии с распределением X^2 с $\nu = (r-1)(c-1)$ степенями свободы.

Критерий Кресси-Рида (К4) имеет вид:

$$CR(\lambda) = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{2}{\lambda(1+\lambda)} A_{ij}\left[\left(A_{ij}/E_{ij}\right)^{\lambda} - 1\right] . \quad (9)$$

Величина λ-параметр алгоритма, в данном случае λ=2/3. Статистика этого критерия имеет распределение X^2 с $\nu = (r-1)(c-1)$ степенями свободы.

Критерий Хеллингера (критерийК5) имеет вид:

$$BWH(\alpha) = \sum_{k=1}^{r}\sum_{l=1}^{c}\left(\frac{A_{ij} - E_{ij}}{\alpha\sqrt{A_{ij}} + (1-\alpha)\sqrt{E_{ij}}}\right), \quad (10)$$

статистика этого критерия также имеет распределение χ^2 с $\nu = (r-1)(c-1)$ степенями свободы при уровне доверительной вероятности α.

Критерий Зелтермана (критерий К6) имеет вид:

$$D_z^2 = X^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} A_{ij}/E_{ij} + rc, \quad (11)$$

где X^2 вычислен по условию (4). Статистика этого критерия, аналогично предыдущим, также имеет распределение χ^2 с $\nu = (r-1)(c-1)$ степенями свободы при уровне доверительной вероятности α.

Полученные результаты

Приведенные в табл. 7 –10 данные проанализированы с использованием критериев К1…К6. Результаты этого анализа приведены в табл. 12 – 15.

При анализе данных, приведённых в табл. 12…табл. 15, отмечено следующее. Значение критерия К1, то есть диагностика Симонова-Цая, на порядок меньше критической величины, равной 0,25. Это означает, что применение всех остальных, использованных в работе критериев, являющихся функцией величины X^2, определяемой по условию (4), допустима и полученные результаты статистически корректны. Анализ результатов применения критериев К2…К6 привёл авторов данной работы к следующим выводам.

Таблица 12

Статистический анализ данных о степени неприятия советской системы разными слоями населения

Результаты анализа	Критерии					
	К1	К2	К3	К4	К5	К6
Численное значение критерия	0,0414	136,9924	134,9924	135,9939	135,8903	136,68340
Величина Pv	-	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Принятая гипотеза	Да	H_0	H_0	H_0	H_0	H_0

Таблица 13

Статистический анализ данных о количественном составе социальных групп, выражавших разную степень враждебности по отношению к режиму

Результаты анализа	Критерии					
	К1	К2	К3	К4	К5	К6
Численное значение критерия	0,0463	34,2433	34,3739	34,2578	34,7734	34,2044
Величина Pv	-	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Принятая гипотеза	Да	H_0	H_0	H_0	H_0	H_0

Таблица 14

Статистический анализ данных о количественном составе социальных групп, выражавших своё неприятие по причине арестов

Результаты анализа	Критерии					
	К1	К2	К3	К4	К5	К6
Численное значение критерия	0,0581	48,2744	50,2677	48,7308	54,4550	48,3432
Величина Pv	-	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Принятая гипотеза	Да	H_0	H_0	H_0	H_0	H_0

Таблица 15

Статистический анализ о национальном составе лиц, заявивших о неприятие советской власти

Результаты анализа	Критерии					
	К1	К2	К3	К4	К5	К6
Численное значение критерия	0,0421	197,3707	194,8412	194,8418	192,238	197,3514
Величина Pv	-	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Принятая гипотеза	Да	H_0	H_0	H_0	H_0	H_0

Применение критерия X^2 даёт результаты, которые численно близки к результатам применения критериев К3…К6, которые не добавляют существенно новой информации. Этот вывод подтверждается тем, что величина доверительной вероятности Pv для всех упомянутых критериев практически одинакова. Результаты применения критериев К2…К6 дают возможность принять статистическую гипотезу H_0 о том, что отсутствует устойчивая статистическая связь между фактором строк и факторов столбцов для табл. 7 – 10.

При оценке содержательного смысла полученных результатов следует исходить из того, что опрос проводили среди перемещённых лиц, то есть среди людей, уже сделавших определённый выбор и являющимися, в какой- то мере, единомышленниками. Это позволяет сформулировать такую нестатистическую гипотезу, подкреплённую результата-

ми оценки статистической гипотезы. Степень неприятия советской власти перемещёнными лицами не зависела от их национального состава, уровня квалификации, принадлежности к одной из социальных групп, особенностей личной биографии в советский период. Более подробный анализ этого результата, относящийся к сфере истории и социологии, выходит за пределы компетенции авторов.

Выводы

1. Для анализа многоклеточных таблиц (категоризированных данных) предложена двухступенчатая процедура, отличающаяся тем, что на первой ступени использована диагностика Симонова-Цая для оценки возможности последующего анализа, основанного на использовании величины X^2.

2. Сравнение полученных результатов проверки статистической гипотезы об отсутствии зависимости между фактором строк и фактором столбцов в многоклеточной таблице показало, что критерии её проверки X^2, отношения правдоподобия, Кресси-Рида, Хеллингера и Зелтермана дают близкие результаты.

3. Для вычислений использованы данные анкетирования перемещённых лиц, полученные в результате реализации гарвардского проекта в 1948-1951 г.г.

4. Установлено, что степень неприятия советской власти перемещёнными лицами не зависела от их национального состава, уровня квалификации, принадлежности к одной из социальных групп, особенностей личной биографии в советский период.

Список литературы

1. *Сидоренко Е.В. Методы математической обработки в психологии / Е.В. Сидоренко. – СПб.: ООО «Речь», 2000. – 350 с.*
2. *Лапач С.Н. Статистика в науке и бизнесе / С.Н. Лапач, А.В. Чубенко, П.Н. Бабич. – К.: Изд. «МОРИОН», 2002. – 640 с.*
3. *Медведев А. Гарвардский проект: полвека спустя http: / А. Медведев [Электронный ресурс]. – Режим доступа: http://www.rg.ru/prilog/union/03-09-04/11.shtm/-21.07.2014 г. – Загл. с экрана.*
4. *Кодин Е.В. Гарвардский проект / Е.В. Кодин. – М.: «Российская политическая энциклопедия» (РОССПЭН), 2003. – 208 с.*
5. *Кендалл М. Статистические выводы и связи / М. Кендалл, А. Стьюарт. – М.: Наука, 1973. – 895 с.*
6. *Закс Л. Статистическое оценивание / Л. Закс. – М.: Статистика, 1976. – 598 с.*
7. *Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. – М.: ФИЗМАТЛИТ, 2006. – 816 с.*
8. *Гайдышев И.П. Моделирование стохастических и детерминированных систем: Руководство пользователя программы AtteStat / И.П. Гайдышев. – Курган: Б.И., 2013. – 496 с.*
9. *Авторханов А. Загадка смерти Сталина / А. Авторханов. – М.: Слово, 1992. – 142 с.*

Поступила в редколлегию 8.09.2014

Рецензент: д-р техн. наук, проф. В.А. Гороховатский, Харьковский институт банковского дела Университета банковского дела НБУ, Харьков.

СТАТИСТИЧНИЙ АНАЛІЗ ПРИЧИН АНТИРАДЯНСЬКИХ НАСТРОЇВ У ПЕРЕМІЩЕНИХ ОСІБ ЗА ДАНИМИ ГАРВАРДСЬКОГО ПРОЕКТУ

В.Ю. Дубницький, А.І. Ходирєв

Для аналізу багатоклітинних таблиць (категоризованих даних) запропоновано двоступеневу процедуру, яка відрізняється тим, що на першій ступені використано діагностику Симонова-Цая для оцінки можливості подальшого аналізу, заснованого на використанні величини X^2. Порівняння отриманих результатів перевірки статистичної гіпотези про відсутність залежності між чинником рядків і чинником стовпців в багатоклітинній таблиці показало, що критерії її перевірки X^2, відношення правдоподібності, Кресі-Ріда, Хеллінгера та Зелтермана дають близькі результати. Для обчислень використані дані анкетування переміщених осіб, отримані в результаті реалізації гарвардського проекту в 1948-1951 р.р. Встановлено, що степінь неприйняття радянської влади переміщеними особами не залежала від їх національного складу, рівня кваліфікації, приналежності до однієї з соціальних груп, особливостей біографії в радянський період.

Ключові слова: *категоризовані дані, багатоклітинні таблиці, перевірка статистичних гіпотез, діагностика Симонова-Цая, критерій хі-квадрат, критерій відношення правдоподібності, критерій Кресі-Ріда, критерій Хеллінгера, критерій Зелтермана.*

STATISTICAL ANALYSIS OF THE REASONS OF ANTI-SOVIET MOOD IN DISPLACED PERSONS UNDER HARVARD PROJECT DATA

V.Yu. Dubnitskiy, A.I. Khodyrev

To analyze multi-cell tables (categorized data) we proposed a two-stage procedure characterized in application of Simonov-Tsai diagnostic at the first stage to evaluate the opportunity of subsequent analysis based on Pearson chi-square test. Comparing the obtained results of verification of statistical hypothesis of absent dependence between liner factor and column factor in a multi-cell table we found that criteria used for its verification: Pearson chi-square test, plausibility relation, power-divergence family Cressie-Read, blended weight Hellinger and Zelterman's statistic give close results. For calculation we used the data of DP questioning obtained from implementation of Harvard Project in 1948-1951. It was stated that the degree of disapproval of Soviet regime did not depend on their ethnic origin, qualification level, belonging to a certain social group, individual life circumstances in Soviet period.

Keywords: *categorized data, multi-cell tables, verification of statistical hypotheses, Simonov-Tsai diagnostic, Pearson chi-square test, plausibility relation criterion, power-divergence family Cressie-Read, blended weight Hellinger and Zelterman's statistic.*