УДК 579.234.3

ПРИМЕНЕНИЕ МЕТОДОВ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ ДЛЯ ОЦЕНКИ РЕЗУЛЬТАТОВ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

А.И. Трубей, Ю.И. Иванченко

В статье рассматривается методика, основанная на статистическом критерии хи-квадрат, как мере статистической значимости связи между различными выборками. Данная методика применялась для оценки статистических данных о численности населения Республики Беларусь. В результате установлены определенные закономерности в распределении населения областей и города Минска (по возрасту и возрастным группам, состоянию в браке, образовательному уровню и удельному весу мужчин и женщин).

Введение

Широкое применение новых информационных технологий, компьютерной техники позволило значительно усовершенствовать методологию и методику социологических исследований, практику сбора и обработки социологических данных, что привело к развитию математических методов социологического анализа [1]. Они позволяют осуществлять анализ и интерпретацию первичной социологической информации, а также верифицировать полученные данные. Предлагается методика оценки статистических данных о численности населения Республики Беларусь с использованием критерия хи-квадрат. В качестве исходного материала для анализа используются официальные статистические данные о численности населения Республики Беларусь, а также его составу по полу, возрасту, состоянию в браке, уровню образования. Реализация данной методики с использованием современных информационных технологий позволит научно обосновать принимаемые решения и усовершенствовать обработку результатов социологических исследований.

1. Теоретические основы критерия хи-квадрат, используемого для проверки однородности статистических выборок

Рассмотрим $M \geq 2$ независимых выборок, каждая из которых есть реализация некоторой полиномиальной схемы с исходами $1, \dots, N$. Объемы выборок равны $t_1, \dots, t_M$. Обозначим $t = \sum_{k=1}^{M} t_k$ и $t_0 = min\{t_1, \dots, t_M\}$; $\nu_{k,i}$ – частота (число появлений) символа i в k-й выборке.

Полиномиальная схема с исходами $1, \dots, N$ определяется вероятностями

$$p_{k,1}, \dots, p_{k,N} > 0; p_{k,1} + \dots + p_{k,N} = 1$$

Полагаем, что выборки статистически однородны и что для них выполняется гипотеза H_0, если

$$p_{k,1} = \dots = p_{k,N}; i = 1, \dots, N;$$

в противном случае будем говорить, что выполняется альтернатива H_1. При альтернативе вероятности исходов могут быть постоянными или изменяться с ростом объемов выборок.

Для проверки однородности полиномиальных выборок обычно используют статистику хи-квадрат [2]

$$\chi^2 = \sum_{k=1}^{M}\sum_{i=1}^{N}\frac{\left(\nu_{k,i}-t_k m_i/t\right)^2}{t_k m_i/t} = t\sum_{k=1}^{M}\sum_{i=1}^{N}\frac{1}{t_k m_i}\left(\nu_{k,i}-m_i\frac{t_k}{t}\right)^2,$$

где $m_i = \sum_{k=1}^{M}\nu_{k,i}$.

При $t_0 \to \infty$ данная статистика сходится по распределению к случайной величине $\chi_{(M-1)(N-1)}$, имеющей распределение хи-квадрат с $(M-1)(N-1)$ степенями свободы [2].

Проверяемая гипотеза принимается, когда вычисленное значение статистики χ^2 не превышает критического значения χ_p^2, или достигнутый уровень значимости (p -value) [3]

$$p = p\left(\chi^2 > \chi_p^2\right) = \frac{1}{2^{(M-1)(N-1)/2}\Gamma\left(\frac{(M-1)(N-1)}{2}\right)}\int_{\chi_p^2}^{\infty} x^{\frac{(M-1)(N-1)}{2}-1}e^{-\frac{x}{2}}dx$$

больше заданного уровня значимости (вероятности ошибки 1-го рода) α, где $\Gamma\left(\frac{(M-1)(N-1)}{2}\right)$ – Гамма-функция Эйлера.

При этом для всех ожидаемых частот должно соблюдаться условие: $t_k p_{k,i} \geq 5$. Рекомендуется установить уровень значимости: $min\left\{1,\frac{5}{M}\right\}$.

2. Методика оценки однородности статистических выборок с использованием критерия хи-квадрат

Методика основана на выявлении различия дискретных вероятностных распределений по наблюдаемым выборкам и состоит из следующих этапов:

1. Формирование статистических выборок $\nu_{k,i}$, упорядоченных по определенным признакам, например, данные о количественном и качественном составе населения, полученные в результате переписи.

2. Вычисление статистики хи-квадрат однородности.

3. Вынесение решения по правилу: принимается гипотеза H_0, если $\chi^2 \leq \chi_p^2$; H_1, если $\chi^2 > \chi_p^2$.

4. Если $\chi^2 \leq \chi_p^2$, то полагаем, что все выборки однородны.

5. Если $\chi^2 > \chi_p^2$, это означает, что в комплекте не все выборки однородны. Для выявления неоднородных выборок осуществим C_M^2 попарных сравнений выборок по статистике χ^2 однородности, вычисленной для $M = 2$. В каждом из этих случаев мы будем иметь 2 независимые выборки, $N-1$ степеней свободы и формула примет вид [3]:

$$\chi_{k,l}^2 = t_k t_l \sum_{i=1}^{N}\frac{1}{\nu_{k,i}+\nu_{l,i}}\left(\frac{\nu_{k,i}}{t_k}-\frac{\nu_{l,i}}{t_l}\right)^2,$$

где $\nu_{k,1}, \dots, \nu_{k,N}$; $\nu_{l,1}, \dots, \nu_{l,N}$ – наборы частот соответственно k-й и l-й выборок;

$t_k = \sum_{i=1}^{N} \nu_{k,i}$; $t_l = \sum_{i=1}^{N} \nu_{l,i}$.

В результате построим треугольную таблицу размером $(M-1)(N-1)$ со строками $1, \dots, M-1$ и столбцами $2, \dots, M$. На пересечении строки k и колонки l $(k < l)$ приведено соответствующее значение статистики $\chi^2_{k,l}$, полученное при сравнении k-й и l-й выборок. Если статистика $\chi^2_{k,l}$, для $N-1$ степеней свободы не превышают критическое значение $\chi^2_{N-1,p}$, которое соответствует уровню значимости p, то полагаем, что выборки однородны. В противном случае считаем, что выборки неоднородны.

6. В некоторых случаях представляет интерес оценить постоянство вероятности одного конкретного исхода. При $N = 2$ мы имеем M последовательных наблюдений, в каждом из которых некоторый исход l осуществляется соответственно $\nu_1, \nu_2, \dots, \nu_j, \dots, \nu_M$ раз. Проверяется гипотеза о том, что исход l во всех наблюдениях имеет одну и ту же, постоянную, хотя и неизвестную вероятность p. Оценкой для p должна служить частота исхода l по всей совокупности данных: $p^* = 1 - q^* = \frac{1}{t}\sum_{j=1}^{M} \nu_j$ и формула для статистики хи-квадрат однородности примет вид [3]:

$$\chi^2 = \sum_{j=1}^{M} \frac{\left(\nu_j - t_j p^*\right)^2}{t_j p^* q^*} = \frac{1}{p^* q^*} \sum_{j=1}^{M} \frac{\nu_j^2}{t_j} - t\frac{p^*}{q^*}$$

с $M-1$ степенями свободы.

Следует отметить, что если выборки «близки» к однородным, то статистика χ^2 будет иметь нецентральное распределение хи-квадрат с тем же числом степеней свободы и ее значение будет пропорционально объему выборки. Использование больших выборок может привести к решению об отклонении гипотезы H_0, хотя неоднородность выборок может быть несущественной и иногда ею можно пренебречь. Кроме того, существующие таблицы нецентрального хи-квадрат распределения недостаточно полны и в статистических приложениях широко используют различные приближения с помощью «центрального» хи-квадрат распределения и нормального распределения. Поэтому при формировании выборок переписи населения, рассматриваемых в разделе 3, в качестве частот исходов будем указывать пропорциональные частоты – целые числа тысяч человек (тыс. чел).

3. Практическое применение методики для оценки результатов переписи населения Республики Беларусь

Основным источником информации о количественном и качественном составе населения являются переписи, которые, как правило, проводятся не реже одного раза в 10 лет. Многообразие количественных данных, полученных в результате переписей и оценок, требует соответствующей обработки и анализа. Применение современных информационных технологий для их обработки нацелено на получение надежной информации о различных аспектах демографического поведения. С целью выявления закономерностей в статистических данных переписи 2009 года [4], упорядоченных по

соответствующим признакам (административно-территориальному делению, полу, возрасту, возрастным группам, состоянию в браке, уровню образования и т. д.), а также переписей и оценок за другие годы [5,6] используем приведенную в разделе 2 методику.

1. Распределение населения Республики Беларусь по возрасту и возрастным группам. В таблице 1 приведено распределение населения областей и города Минска (далее – регионов) по возрасту (по данным переписи 2009 года) [4]. Колонки таблицы являются независимыми полиномиальными выборками с $N = 8$ исходами.

Таблица 1

Распределение населения регионов по возрасту (тыс. чел.)

Числ. населения в возрасте, лет:	Бр. обл.	Вит. обл.	Гом. обл.	Гр. обл.	Мин. обл.	Мог. обл.	г. Минск
0 – 9	152	109	146	107	140	107	174
10 – 19	171	136	168	126	159	128	214
20 – 29	205	185	224	157	212	171	370
30 – 39	199	170	205	148	195	157	270
40 – 49	210	190	218	171	223	165	258
50 – 59	195	185	209	143	206	165	264
60 – 69	120	110	114	95	124	88	145
70 и старше	149	146	157	125	164	120	143

Проверим гипотезу об однородности выборок для всех регионов. С этой целью осуществим C_7^2 попарных сравнений колонок (выборок) таблицы по статистике χ^2, вычисленной для $M = 2$. В результате построим треугольную таблицу, размером 6×6 (таблица 2).

Таблица 2

Попарные сравнения распределения населения регионов по возрасту

Регионы	Вит. обл.	Гом. обл.	Гр. обл.	Мин. обл.	Мог. обл.	г. Минск
Бр. обл.	5,15	1,51	1,76	2,39	2,05	**23,63**
Вит. обл.		3,09	2,35	0,96	1,98	**26,75**
Гом. обл.			2,28	1,49	***0,22***	**19,10**
Гр. обл.				**0,84**	2,54	**25,85**
Мин. обл.					1,32	**27,19**
Мог. обл.						**16,15**

При сравнении выборки населения по возрасту города Минска с аналогичными выборками по областям все значения статистики χ^2 превышают критическое значение $\chi^2_{7,p} = 14{,}07$, которое соответствует уровню значимости $p = 0{,}05$. Это означает, что отклонение от гипотезы об однородности распределений весьма значимо. Следовательно, распределение населения города Минска по возрасту существенно отличается от распределения населения по областям.

В таблице 3 приведено распределение населения областей и города Минска по возрастным группам (по данным переписи 2009 года). Строки таблицы являются независимыми полиномиальными выборками с $N = 3$ исходами.

Таблица 3

Распределение населения регионов по возрастным группам (тыс. чел.)

Регионы	Моложе трудоспособного возраста	В трудоспособном возрасте	Старше трудоспособного возраста
Брестская обл.	249	836	316
Витебская обл.	181	747	303
Гомельская обл.	237	881	323
Гродненская обл.	176	642	254
Минская обл.	229	857	336
Могилевская обл.	176	675	248
г. Минск	263	1 214	359

Результаты попарных сравнений строк таблицы по статистике χ^2 однородности, вычисленной для $M = 2$, представлены в таблице 4.

Таблица 4

Попарные сравнения распределения населения регионов по возрастным группам

Регионы	Вит. обл.	Гом. обл.	Гр. обл.	Мин. обл.	Мог. обл.	г. Минск
Бр. обл.	5,07	0,99	0,99	1,55	1,43	**14,63**
Вит. обл.		2,68	1,35	1,12	1,73	**12,23**
Гом. обл.			0,60	0,60	***0,09***	**8,70**
Гр. обл.				**0,05**	0,56	**11,59**
Мин. обл.					0,44	**12,24**
Мог. обл.						**6,69**

При сравнении выборки города Минска с выборками по областям, было установлено, что все значения статистики χ^2 превышают критическое значение $\chi^2_{2,p} = 5{,}99$ для уровня значимости $p = 0{,}05$. Следовательно, распределение населения Минска по возрастным группам, как и по возрасту, существенно отличается от распределений по областям. При сравнениях по возрастным группам населения Гродненской и Минской областей, а также Гомельской и Могилевской областей получим значения статистики χ^2, которые соответствуют $p > 0{,}95$. Это свидетельствует об однородности распределения данных пар областей по указанному признаку.

2. Распределение женщин Республики Беларусь по состоянию в браке. В таблице 5 приведено распределение женщин областей и города Минска по состоянию в браке (по данным переписи 2009 года). Строки таблицы являются независимыми полиномиальными выборками с $N = 4$ исходами.

Таблица 5

Распределение женщин в возрасте 15 лет и старше по состоянию в браке (тыс. чел.)

Регионы	состоящие в браке	никогда не состоявшие в браке	вдовые	разведенные
Брестская обл.	350	103	120	57
Витебская обл.	298	97	116	65
Гомельская обл.	348	117	124	75
Гродненская обл.	267	80	102	43
Минская обл.	353	100	136	61
Могилевская обл.	264	87	96	56
г. Минск	426	210	116	108

Результаты попарных сравнений строк таблицы по статистике χ^2 однородности, вычисленной для $M = 2$, представлены в таблице 6.

Таблица 6

Попарные сравнения распределения женщин по состоянию в браке

Регионы	Вит. обл.	Гом. обл.	Гр. обл.	Мин. обл.	Мог. обл.	г. Минск
Бр. обл.	2,53	2,53	**0,51**	**0,88**	1,86	**24,94**
Вит. обл.		***0,48***	2,12	1,95	***0,22***	**19,51**
Гом. обл.			3,91	3,21	***0,05***	**15,58**
Гр. обл.				**0,27**	2,08	**25,35**
Мин. обл.					2,15	**31,64**
Мог. обл.						**14,86**

При сравнении выборки женщин города Минска с аналогичными выборками по областям все значения статистики χ^2 превышает критическое значение $\chi^2_{3,p} = 11{,}34$, которое соответствует уровню значимости $p = 0{,}01$. Следовательно, распределение женщин города Минска по состоянию в браке существенно отличается от соответствующих распределений по областям. В то же время, при сравнениях выборок женщин трех восточных областей (Витебской, Гомельской и Могилевской) получим значения статистики χ^2, которые соответствуют $p > 0{,}85$. Это означает, что распределения женщин данных областей по состоянию в браке практически не отличаются. Аналогичный результат получим при сравнениях выборок женщин трех западных областей (Брестской, Гродненской и Минской).

При сравнении суммарных выборок первой и второй групп областей получим $\chi^2 = 6{,}36$, которое превышает критическое значение $\chi^2_{3,p} = 6{,}25$, соответствующее уровню значимости $p = 0{,}1$. Это означает, что распределения женщин восточных и западных областей Республики Беларусь по состоянию в браке различаются.

3. Распределение населения Республики Беларусь по образовательному уровню. В таблице 7 приведено распределение населения Республики Беларусь по уровню образования (по данным переписи 2009 года). Строки таблицы являются независимыми полиномиальными выборками с $N = 6$ исходами.

Таблица 7

Распределение населения регионов по уровню образования (тыс. чел.)

Числ. населения в возрасте 10 лет и старше	Из него с уровнем образования					
	высш.	средн. спец.	проф.-технич.	общим средним	общим базовым	общим начальным
Брестская обл.	188	316	126	300	120	168
Витебская обл.	171	303	137	235	118	130
Гомельская обл.	196	330	152	303	128	155
Гродненская обл.	153	271	92	199	94	131
Минская обл.	184	322	152	301	136	161
Могилевская обл.	146	271	128	211	102	113
г. Минск	493	377	112	366	92	97

Результаты попарных сравнений строк таблицы по статистике χ^2 однородности, вычисленной для $M = 2$, представлены в таблице 8.

Таблица 8

Попарные сравнения распределения населения регионов по уровню образования

Регионы	Вит. обл.	Гом. обл.	Гр. обл.	Мин. обл.	Мог. обл.	г. Минск
Бр. обл.	7,67	2,85	4,79	3,10	8,71	**136,95**
Вит. обл.		2,58	5,71	3,86	***0,38***	**132,45**
Гом. обл.			7,14	**0,81**	2,78	**137,08**
Гр. обл.				8,25	7,56	**115,52**
Мин. обл.					3,49	**155,42**
Мог. обл.						**129,05**

В ходе анализа результатов попарных сравнений выборки населения города Минска по образовательному уровню с выборками по областям отметим, что значение статистики χ^2 более чем на порядок превышает значения статистики χ^2 соответствующих сравнений для областей. Это означает, что распределение населения города Минска по образовательному уровню существенно отличается от распределения населения областей. Кроме того, следует отметить схожесть по образовательному уровню населения Витебской и Могилевской областей, а также Гомельской и Минской областей.

4. В таблице 9 приведено распределение населения районов города Минска по уровню образования (по данным переписи 2009 года). Строки таблицы являются выборками с $N = 6$ исходами.

Таблица 9

Распределение населения районов г. Минска по уровню образования (тыс. чел.)

Числ. населения в возрасте 10 лет и старше	Из него с уровнем образования					
	высшим	средн. спец.	проф.-технич.	общим средним	общим базовым	общим начальным
Зав. р-н	47	56	21	53	15	14
Лен. р-н	55	45	15	44	11	12
Моск. р-н	68	59	17	52	12	14
Окт. р-н	42	33	10	32	7	7
Парт. р-н	26	19	6	20	6	6
Перв. р-н	67	40	10	44	9	9
Сов. р-н	53	27	6	34	8	8
Фрун. р-н	98	77	23	67	19	21
Цент. р-н	37	20	5	21	5	5

Результаты попарных сравнений строк таблицы по статистике χ^2 однородности, вычисленной для $M = 2$, представлены в таблице 10.

Таблица 10

Попарные сравнения населения районов г. Минска по уровню образования

Районы	Лен. р.	Мос. р.	Окт. р.	Парт. р.	Перв. р.	Сов. р.	Фрун. р.	Цент. р.
Зав. р.	**2,96**	**4,08**	**4,07**	**2,75**	**11,66**	**12,96**	**5,89**	**9,90**
Лен. р.		***0,29***	***0,39***	***0,33***	3,08	4,39	***0,48***	2,92
Мос. р.			***0,28***	***0,78***	3,08	4,85	***0,52***	2,84
Окт. р.				***0,73***	1,40	2,98	***0,71***	1,72
Парт. р.					1,82	2,12	***0,41***	1,68
Перв. р.						**0,70**	3,20	**0,24**
Сов. р.							4,39	**0,37**
Фрун. р.								2,46

В ходе анализа результатов попарных сравнений выборки по уровню образования населения Заводского района с выборками населения других районов можно установить, что значения статистики χ^2 для пяти степеней свободы значительно превышает значения соответствующих сравнений для других районов. Это означает, что распределение населения Заводского района по уровню образования существенно отличается от распределения населения этих районов по данному признаку.

Действительно, в отличие от других районов, по уровню образования в Заводском районе превалируют лица со средним специальным и профессионально-техническим образованием – 35,4% от населения в возрасте 10 лет и старше. Доля лиц с высшим образованием составляет 21,6%.

С другой стороны, можно выделить две группы районов, у которых при попарных сравнениях значения статистик χ^2 принимают значения, которые соответствуют $p > 0{,}95$, что свидетельствует об однородности соответствующих выборок. В первую группу входят Первомайский, Советский и Центральный районы, во вторую – Ленинский, Московский, Октябрьский, Партизанский и Фрунзенский районы.

При сравнении суммарных выборок населения первой и второй групп получим $\chi^2 = 9{,}63$, что свидетельствует о различии распределение населения соответствующих групп районов по уровню образования. При сравнении выборки Заводского района с суммарными выборками первой и второй групп получим соответственно $\chi^2_{\text{зав},1} = 19{,}62$, $\chi^2_{\text{зав},2} = 6{,}54$. Это согласуется с гипотезой о неоднородности данных выборок.

5. Соотношение мужчин и женщин в Республики Беларусь. В таблице 11 приведено количество мужчин и женщин, проживающих в стране (по материалам переписей и оценок). Строки таблицы являются полиномиальными выборками с $M = 7$ исходами, отражающими динамику численности мужчин, женщин и всего населения в период с 1959 по 2015 годы [5,6].

Таблица 11

Распределение мужчин и женщин страны по материалам переписей и оценок (тыс. чел.)

Годы	1959	1970	1979	1989	1999	2009	2015
Мужчины	**3 570**	4 129	4 421	4 749	4 718	4 420	4 409
Женщины	**4 462**	4 863	5 111	5 403	5 328	5 084	5 072
Всего	**8 032**	8 992	9 532	10 152	10 046	9 504	9 481

Проверим гипотезу о том, что удельный вес мужчин в структуре населения за этот период времени оставался неизменным. Вычислим значение статистики χ^2 однородности для $N = 2$. Получим $\chi^2 = 14{,}65$ с шестью степенями свободы, этому соответствует $p = 0{,}025 < 0{,}05$, и можно считать, что гипотеза о низменности удельного веса мужчин не выполняется при заданном уровне значимости. Однако, если мы не будем принимать в учет результаты послевоенной переписи населения 1959 года, то получим $\chi^2 = 2{,}45$ для пяти степеней свободы, что соответствует $p = 0{,}75$ и можно считать, что указанная гипотеза справедлива. Оценкой для удельного веса мужчин является:

$$p_{\text{муж}} = \frac{30416 - 3570}{65739 - 8032} = \frac{22437}{48496} = 0{,}465212$$

Действительно, по переписи 1959 г. на 1000 мужчин приходилось 1249 женщин, по переписи 1989 г. – 1138 женщин, по переписи 1999 г. – 1129 женщин, по переписи 2009 года – 1150 женщин. Нарушения структуры населения Беларуси по полу, образовавшиеся в годы Великой Отечественной войны, к настоящему времени в значительной степени сгладились и дают о себе знать только в возрастах старше 80 лет.

Заключение

На основании анализа результатов переписей и оценок населения Республики Беларусь с использованием критерия хи-квадрат установлены определенные закономерности в распределении населения областей и города Минска (по возрасту и возрастным группам, состоянию в браке, образовательному уровню и удельному весу мужчин и женщин).

Применение современных информационных технологий и математических методов позволяет автоматизировать обработку социологических исследований, нацеленных на получение достоверной информации по важнейшим демографическим показателям Республики Беларусь.

Список литературы

1. Данилов, А.Н. Белорусская социология сегодня: проблемное поле и истоки оптимизма / А.Н. Данилов // Социологические исследования. СОЦИС, – 2014. – № 8. – С. 21–31.

2. Зубков, А.М. Об одной статистике для проверки однородности полиномиальных выборок / А.М. Зубков, Б.И. Селиванов // Дискретная математика. – 2014.– т. 26, вып. 3 – С. 30–44.

3. Крамер, Г. Математические методы статистики / Г. Крамер. – М.: Мир, 1976. – 648 с.

4. Перепись населения 2009 года [Электронный ресурс]. — 2016. — Режим доступа: http://belstat.gov.by/.— Дата доступа: 15.02.2016.

5. Численность мужчин и женщин Республики Беларусь [Электронный ресурс]. — 2016. — Режим доступа: http://belstat.gov.by/.— Дата доступа: 15.02.2016.

6. Демографический ежегодник Республики Беларусь, 2015 [Электронный ресурс]. —2016. — Режим доступа: http://belstat.gov.by/. — Дата доступа: 15.02.2016.

Трубей Антон Иванович, старший научный сотрудник НИИ прикладных проблем математики и информатики Белорусского государственного университета, trubeia@mail.ru

Иванченко Юрий Иванович, заведующий лабораторией НИИ прикладных проблем математики и информатики Белорусского государственного университета, кандидат технических наук, IvanchenkoYI@bsu.by