СТАТИСТИКА И ЭКОНОМИЧЕСКОЕ ИЗМЕРЕНИЕ

УДК 311.21

О ПРОВЕРКЕ ФАКТИЧЕСКОЙ РЕПРЕЗЕНТАТИВНОСТИ СОЦИАЛЬНОЙ ВЫБОРКИ

В.В. Глинский, Ю.В. Гусев, Н.И. Овечкина, Е.С. Шмарихина
Новосибирский государственный университет
экономики и управления «НИНХ»
E-mail: s444@ngs.ru

В статье рассматриваются вопросы, связанные с проверкой представительности результатов выборочных социологических, демографических обследований; особенностью генеральных совокупностей в таких исследованиях (население в целом, его отдельные типы, классы) является то обстоятельство, что в большинстве случаев априори неизвестны значения параметров этих множеств, что затрудняет применение традиционных подходов к оценке репрезентативности полученных выборок, основанных на ЦПТ (центральной предельной теореме); в работе сформулированы и апробированы на реальных информационных массивах дополнительные условия (проверка фактической информативности типообразующих признаков) , позволившие снизить риски смещения результатов социальных выборок.

Ключевые слова: представительность, параметр, статистика, выборка, информативность, дендрит, значимые связи.

ON CHECKING OF THE REAL REPRESENTATIVENESS OF SOCIAL SAMPLING

V.V. Glinskiy, Yu.V. Gusev, N.I. Ovechkina, E.S. Shmarikhina
Novosibirsk State University of Economics and Managment
E-mail: s444@ngs.ru

The article considers the issues relating to checking of representativity of results of sampling sociological, demographic observations. The entire assembly feature in such researches (general population, its individual types, classes) is that in most cases parameters values of these multitudes are a priori unknown which makes it difficult to apply traditional approaches to the appraisal of representativeness of result samplings based on CLT (central limit theorem). The paper formulates and evaluates on real information arrays additional conditions (checking of real informativeness of the type-forming attributes), allowing to decrease risks of human bias.

Key words: representativity, parameter, statistics, sampling, informativeness, dendrite, significant relations.

Данные, полученные в результате выборочного обследования различных категорий населения, являются конкретными реализациями случайных величин (пол, возраст, уровень образования, национальность, длительность заболевания и т. п.), следовательно, всегда существует ненулевая вероятность получения непредставительной выборки, и даже в случае полного соблюдения принципов выборочного обследования – случайности, равной возможности для единиц генеральной совокупности попасть в выборочную [3–5, 8]. Обычно под представительностью понимают адекватность структуры выборочной совокупности структуре генеральной совокупности. Идеальную репрезентативность получают в случае полной адекватности обеих структур, что реализовать практически невозможно, поэтому считают представительной выборку (в экономике, социологии, демографии; в технике, медицине и т.п. критерии надежности, как правило, более жесткие), которая обеспечивает отклонения значений основных характеристик (статистик) выборочной совокупности не более чем на 5% относительно параметров генеральной совокупности. Следовательно, в случае, когда имеются данные по генеральной совокупности, фактическую репрезентативность выборки можно проверить сравнением показателей по генеральной и выборочной совокупностям, и если фактическая ошибка не превышает, например, 5%, можно считать, что фактически полученная выборка представительна в заданных ограничениях.

В практических исследованиях, особенно это касается социологических, демографических обследований, чаще встречаются ситуации, когда нет показателей по генеральной совокупности (собственно, выборка обычно и организуется для того, чтобы получить оценки параметров генеральной совокупности). В таких ситуациях рекомендуется организовать две независимые выборки и затем сравнить их параметры. Этот прием в теоретическом аспекте достаточно привлекателен и корректен, однако в прикладном смысле практически не применим. Здесь следует учитывать два существенных обстоятельства: во-первых, даже одно обследование провести достаточно сложно и недешево, а во-вторых, если результаты этих независимых выборок не совпадут, то возникает вопрос, какая из них непредставительна. Учитывая данные соображения, мы предлагаем определить представительность выборки проверкой фактической информативности априорно информативных признаков. Действительно, если для совокупности, например, больных получилось, что длительность и вид болезни не связаны с системой остальных признаков программы наблюдения, значит, реализована плохая выборка и продолжать дальнейший анализ не имеет смысла.

Известны два подхода к определению информативности признака в системе. Так, предлагается оценивать информативность величиной $\sum_{i=1}^{n} \Gamma_{ij}$, где $\Gamma_{YX} = I(YX)/H(X)$ – показатель влияния X на Y; $H(X) = -\sum P(X_j) \cdot \log_2 P(X_j)$ – неопределенность (энтропия) случайной величины X; $I(XY) = H(Y) - H_X(Y)$ – снятая неопределенность (информация), где $H_X(Y) = -\sum_j P(X_j) \sum_i^X P(X_j Y_j) \cdot \log_2 P(X_j, Y_i)$ – средняя условная энтропия случайной величины Y при условии X; $H(Y) = -\sum P(Y_i) \cdot \log_2 P(Y_i)$ – энтропия случайной величины Y [6, с. 128].

Наиболее информативным будет, следовательно, признак, имеющий $\max\sum\Gamma_{ij}$. Ввиду того, что Γ_{YX} – показатель влияния X на Y, $\sum\Gamma_{ij}$ будет давать, с нашей точки зрения, оценку информативности признака Y как результата. В сложной системе взаимосвязей признак может одновременно выступать и как фактор, и как результат, поэтому $\sum\Gamma_{ij}$ не в полной мере будет характеризовать его информативность.

В соответствии с другим подходом [7, с. 128–130] информативность признака определяется суммой коэффициентов взаимной информации либо иных показателей связи (коэффициентов парной корреляции – по модулю, коэффициентов взаимной сопряженности и т.п.). Этот подход, с нашей точки зрения, лучше характеризует информативность признаков системы, однако следует указать на недостаток, присущий обоим подходам. Предположим, что в системе из 30 признаков мы получили по одному из них $\sum_{i=1}^{30} R(X_1X_i) = 10,510$, а по второму – $\sum_{i=1}^{30} R(X_2X_i) = 10,508$ (R – показатель связи). Имеется ли основание называть X_1 более информативным, чем X_2? Наверное, нет. Значения коэффициентов являются конкретными реализациями вероятностного процесса, взаимодействием необходимости и случайности, и в данном случае разница носит, скорее, случайный характер.

С учетом этого обстоятельства определению информативности признака, когда наиболее информативным будет более влиятельный признак, в значительной мере отвечает дискретный показатель. Мы предлагаем два таких показателя: число значимых связей для данного признака в системе и число связей в дендрите, построенном для данной системы признаков. Эти показатели в лучшей мере будут характеризовать связанность признака, его информативность.

Рассчитаем показатели информативности для ситуации проверки репрезентативности выборочной совокупности, сформированной из контингента больных работников одного из заводов г. Новосибирска. Цель исследования – изучение факторов, определяющих уровень гипертонической болезни на данном предприятии. В качестве исходной расчетной базы используем матрицу нормированных коэффициентов сопряженности Пирсона, рассчитанную для выборочной совокупности (табл. 1).

$$C' = \frac{C}{C_{\max}}; \quad C = \sqrt{\frac{\chi^2}{\chi^2 + n}}; \quad C_{\max} = \sqrt{\frac{\min\{n-1; m-1\}}{\min\{n-1; m-1\}+1}},$$

где C' – нормированный коэффициент взаимной сопряженности; C – коэффициент сопряженности Пирсона.

Для каждого признака рассчитаем:

1) $\sum_{i=1}^{11} C'_{ij}$ при $i \neq j$, где C'_{ij} – коэффициент связи между X_i и X_j;

2) N_i – число значимых по критерию χ^2 связей.

Результаты расчетов приведены в табл. 2.

Таблица 1

Матрица коэффициентов взаимной сопряженности

Признак	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Y – длительность заболевания	1000	362	462	537	415	406	639	492	427	462	427
X_1 – пол	362	1000	141	385	198	837	769	622	089	308	308
X_2 – возраст	462	141	1000	907	172	109	197	491	495	337	385
X_3 – трудовой стаж	537	385	907	1000	510	387	331	426	553	337	362
X_4 – качество вентиляции	415	198	172	510	1000	198	267	370	141	406	141
X_5 – курение	406	837	109	387	198	1000	822	622	000	337	242
X_6 – употребление алкоголя	639	769	197	331	267	822	1000	471	198	000	000
X_7 – увлечения	492	622	491	426	370	622	471	1000	385	000	308
X_8 – наследственность	427	089	495	553	141	000	198	385	1000	198	362
X_9 – характер труда	462	308	337	337	406	337	000	000	198	1000	308
X_{10} – сверхурочные работы	427	308	385	362	141	242	000	308	362	308	1000

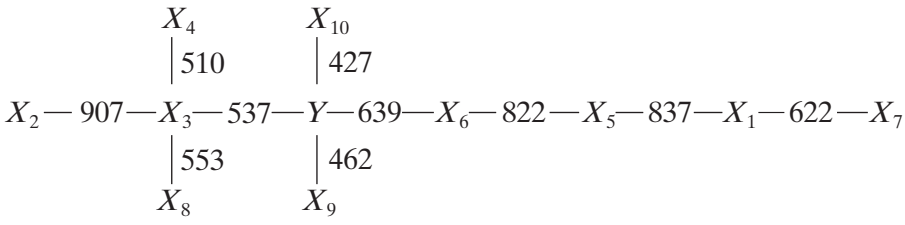
П р и м е ч а н и е. Большинство признаков программы атрибутивные, следовательно, адекватной им мерой связи являются коэффициенты сопряженности, теоретико-информационные меры и т. п. Коэффициенты умножены на 1000, подчеркнуты значимые по критерию χ^2 показатели.

Таблица 2

Оценка информативности признаков программы наблюдения

Признак	C′		Число значимых связей		Число связей в дендрите	
	$\sum C'_{ij}$	ранг	N_i	ранг	M_i	ранг
Y – длительность заболевания	4629	2	9	1	4	1,5
X_1 – пол	4019	4	4	7,5	2	4
X_2 – возраст	3696	6	5	4,5	1	8,5
X_3 – трудовой стаж	4735	1	7	3	4	1,5
X_4 – качество вентиляции	2818	10	4	7,5	1	8,5
X_5 – курение	3960	5	5	4,5	2	4
X_6 – употребление алкоголя	3694	7	4	7,5	2	4
X_7 – увлечения	4187	3	8	2	1	8,5
X_8 – наследственность	2848	8	4	7,5	1	8,5
X_9 – характер труда	2693	11	2	10,5	1	8,5
X_{10} – сверхурочные работы	2843	9	2	10,5	1	8,5

Более наглядным средством оценки информативности является анализ связей в дендрите. Построим дендрит, используя данные табл. 2 и схему дендрита на коэффициентах взаимной сопряженности:



В дендрите отчетливо видны наиболее информативные признаки. Большим числом характеризуются X_3 (трудовой стаж) и Y (длительность заболевания).

Аналогичные расчеты, только по более широкой программе, проведены по проверке представительности выборки, образующей совокупность хронических больных на том же предприятии. Данные представлены в табл. 3 и на рисунке. Информационность измерена для 33 признаков (в табл. 3 приведены 18 наиболее информативных признаков).

Были установлены следующие показатели для определения информативности:

$$1) \sum_{\substack{i=1 \\ i \neq j}}^{33} \Gamma_{ij}, \; i \neq j; \quad 2) \sum_{\substack{i=1 \\ i \neq j}}^{33} R_{ij},$$

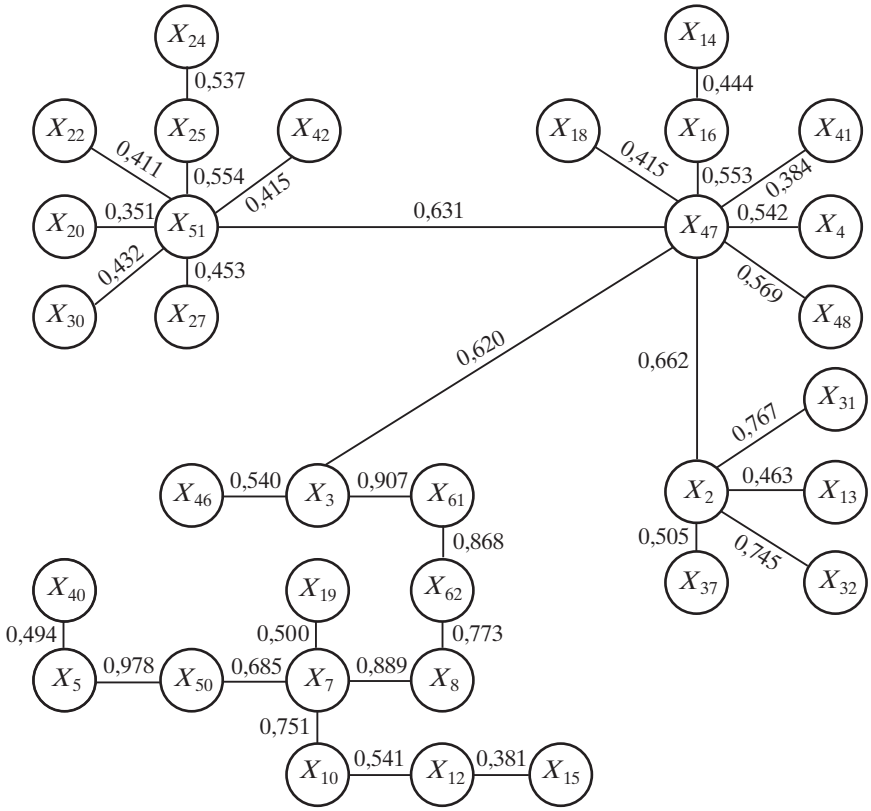
где $R_{XY} = I(XY)/H(XY)$ – коэффициент взаимной информации, предложенный К. Райским;

3) N_i – число значимых по критерию χ^2 связей.

Таблица 3

Оценка информативности признаков программы наблюдения

№	Признак	Γ_{YX}		R_{XY}		Число значимых связей	
		$\sum \Gamma_{YX}$	ранг	$\sum R_{XY}$	ранг	N_i	ранг
1	X_2 – пол	0,853	18	7,84	17	15	10
2	X_3 – возраст	1,797	6	11,762	5	21	6
3	X_5 – профессия	1,350	16	10,237	7	22	5
4	X_{61} – стаж трудовой общий	1,857	3	11,531	6	25	4
5	X_{62} – стаж работы на заводе	1,572	15	10,110	8	18	8
6	X_7 – производственные вредности	2,026	1	12,803	3	26	3
7	X_8 – стаж работы с производственными вредностями	1,889	2	12,168	4	21	6
8	X_{10} – запыленность	1,655	12	8,260	14	19	7
9	X_{12} – загазованность	1,784	7	7,768	12	21	6
10	X_{18} – температура воздуха	1,685	9	8,906	11	21	6
11	X_{19} – производственный шум	1,630	14	8,152	15	17	9
12	X_{24} – число детей	1,841	5	9,695	9	15	10
13	X_{25} – среднедушевой доход	1,845	4	9,615	10	17	9
14	X_{31} – курение	1,753	8	7,864	16	18	8
15	X_{32} – употребление алкоголя	1,636	13	6,806	18	15	10
16	X_{50} – характер труда	0,909	17	8,750	13	17	9
17	X_{51} – длительность заболевания	1,675	11	13,416	1	27	2
18	X_{47} – вид заболевания	1,682	10	13,038	2	28	1



Дендрит на коэффициентах взаимной информации К. Райского

X_{10} – запыленность; X_{12} – загазованность; X_{13} – характер работы (основная поза); X_{14} – сквозняки; X_{15} – вибрация; X_{16} – температура воздуха; X_{18} – освещенность; X_{19} – производственный шум; X_2 – пол; X_{22} – доставка на работу транспортом (пешком); X_{24} – число детей; X_{25} – среднедушевой доход; X_{27} – качество питания; X_3 – возраст; X_{31} – курение; X_{32} – употребление алкоголя; X_{37} – увлечения; X_{38} – занятия спортом; X_4 – житель коренной (или приезжий); X_{41} – место проживания; X_{42} – бытовой шум; X_{46} – сопутствующие заболевания; X_{47} – вид болезни; X_{49} – сверхурочные работы; X_5 – категория профессии; X_{50} – характер труда; X_{51} – длительность заболевания; X_{61} – общий стаж работы; X_{62} – стаж работы на заводе; X_7 – производственная вредность; X_8 – стаж работы с производственными вредностями

Как показывают данные табл. 3, признаки заболеваемости и в этом случае входят в число наиболее информативных, причем по R_{XY} и числу значимых связей они показывают наибольшую «связность». В дендрите (см. рисунок) отчетливо видны наиболее информативные признаки. Большим числом связей характеризуются признаки X_{47} – вид заболевания (связан с 8 признаками – X_4, X_{16}, X_{18}, X_2, X_{48}, X_{51}, X_{41}, X_3); X_{51} – длительность заболевания (7 связей); X_2 – пол (5 связей); X_7 – работа с производственными вредностями (4 связи); X_3 – возраст (3 связи).

Полученные результаты и представленные ранее несколько различаются, но они согласуются в главном: показатели заболеваемости (длительность и вид болезни) являются одними из наиболее информативных признаков программы наблюдения. Следовательно, гипотеза о репрезентативности фактически полученной выборки не может быть отвергнута, а информация, полученная в ходе обследования, может быть применена в дальнейшем анализе.

Таким образом, прежде чем приступать к применению результатов выборочного обследования, необходимо провести проверку репрезентативности выборки. Очевидно, что даже корректно организованное выборочное наблюдение с ненулевой вероятностью может привести к смещенным результатам – в силу действия ЦПТ, также очевидно, что достаточных условий избежать случайной ошибки выборки нет (если не рассматривать ситуацию, когда численность выборки равна численности генеральной совокупности), однако обеспечить выполнение корректных дополнительных необходимых условий вполне возможно, причем это позволит снизить риски принятия последующих решений. При отсутствии известных параметров генеральной совокупности проверка может быть выполнена с использованием оценки информативности априорно-информативных признаков. При этом связность признаков (например, длительность и вид заболевания, как в рассмотренном нами примере) является необходимым условием представительности выборочной совокупности.

Литература

1. *Аптон Г.* Анализ таблиц сопряженности. М.: Финансы и статистика, 1982. 160 с.
2. *Глинский В.В.*, *Ионин В.Г.* Статистический анализ. М.: ИНФРА-М, 2002. 241 с.
3. *Глинский В.В.*, *Серга Л.К.* Нестабильные совокупности: концептуальные основы методологии статистического исследования // Вестник НГУЭУ. 2009. № 2. С. 137–142.
4. *Глинский В.В.* Как измерить малый бизнес // Вопросы статистики. 2008. № 7. С. 73–75.
5. *Глинский В.В.* Мифическая статистика малого бизнеса. Проблемы статистического изучения турбулентных совокупностей // ЭКО (Экономика и организация промышленного производства). 2008. № 9. С. 51–62.
6. *Елисеева И.И.*, *Рукавишников О.В.* Группировка, корреляция, распознавание образов. М.: Статистика, 1977. 144 с.
7. *Славко Т.Л.* Математико-статистические методы в исторических исследованиях. М.: Наука, 1981. 158 с.
8. *Шмарихина Е.С.* Комплексный подход к оценке качества выборочного обследования // Вестник НГУЭУ. 2011. № 1. С. 129–137.

Bibliography

1. *Apton G.* Analiz tablic soprjazhennosti. M.: Finansy i statistika, 1982. 160 p.
2. *Glinskij V.V.*, *Ionin V.G.* Statisticheskij analiz. M.: INFRA-M, 2002. 241 p.
3. *Glinskij V.V.*, *Serga L.K.* Nestabil'nye sovokupnosti: konceptual'nye osnovy metodologii statisticheskogo issledovanija // Vestnik NGUJeU. 2009. № 2. P. 137–142.
4. *Glinskij V.V.* Kak izmerit' malyj biznes // Voprosy statistiki. 2008. № 7. P. 73–75.
5. *Glinskij V.V.* Mificheskaja statistika malogo biznesa. Problemy statisticheskogo izuchenija turbulentnyh sovokupnostej // JeKO (jekonomika i organizacija promyshlennogo proizvodstva). 2008. № 9. P. 51– 62.
6. *Eliseeva I.I.*, Rukavishnikov O.V. Gruppirovka, korreljacija, raspoznavanie obrazov. M.: Statistika, 1977. 144 p.
7. *Slavko T.L.* Matematiko-statisticheskie metody v istoricheskih issledovanijah. M.: Nauka, 1981. 158 p.
8. *Shmarihina E.S.* Kompleksnyj podhod k ocenke kachestva vyborochnogo obsledovanija // Vestnik NGUJeU. 2011. № 1. P. 129–137.