ОЩЕ ЕДНА ГЛЕДНА ТОЧКА КЪМ ПРОБЛЕМА ЗА ОТКАЗИТЕ ПРИ СОЦИОЛОГИЧЕСКИ ИЗСЛЕДВАНИЯ

КАЛОЯН ХАРАЛАМПИЕВ

Резюме: В статията се търси отговор на въпроса колко е допустимият дял на отказите при социологическите изследвания. Проблемът се разглежда през парадигмата на бейсовската статистика. Извеждат се две формули – първата, която позволява предварително да се планира допустимия дял на отказите, и втората, която позволява при вече приключило теренно изследване да се изчисли диапазонът на разминаването в оценките между реализираната и планираната извадка.

Напоследък[1], покрай актуалните в момента електорални изследвания, отново придобива актуалност въпросът колко е допустимият дял на отказите при социологическите изследвания[2]. Моят преглед на литературата показа, че няма точен отговор на този въпрос. Публикациите по темата за отказите при социологическите изследвания основно могат да се разделят на две групи – в едната група попадат тези публикации, които разглеждат причините за отказите (Атанасов и кол. 2006: 40-46; Дерменджиева, Цветкова, Милева 2010).

В другата група са публикациите, които разглеждат последиците от отказите, а именно: „ако съществуват систематични отклонения в мненията на отговорилите и неотговорилите, ниският процент на участие очевидно създава значителен риск от *изместване* на резултатите" (Парчев 1998: 113). И още „Възможно е изобщо да няма изместване (ако мненията на отговорилите и неотговорилите по зададените въпроси не се различават), или пък да има

[1] Виж например отговорите на Цветозар Томов в съвместното му интервю с Андрей Райчев в „Гласове" (http://www.glasove.com/erata-borisov--intervyu-s-andrey-raychev-i-tsvetozar-tomov--ii-chast-16287) или коментарите на Стойко Тонев от 29 септември 2011 в „Делниците на един луд" (http://frognews.bg/news_39164/CHetvarti_kilometar/ или http://www.reduta.bg/?p=1440).

[2] В тази статия няма да се спирам на проблема със заместването на отказалите. Професор Венедиков винаги е настоявал, че заместването е недопустимо, но в неговите публикации не успях да открия никъде експлицитно заявено това твърдение. Това което открих е, че трябва „да наблюдаваме точно тези единици, които са избрани чрез лотарията" (Венедиков 1992: 43; Венедиков 1993: 40). Разбира се, от гледна точка на езика на математиката твърдението „недопустимо е заместването на единици" е пряко следствие от твърдението „наблюдават се точно тези единици, които са избрани". От своя страна аз, в една публицистична статия, писана по повод на предходните местни избори през 2007 година, съм показал защо заместването е опасно и че е недопустимо да се прави (Харалампиев 2007: 16). Това че заместването на единици е (масова) практика е въпрос на друг разговор.

значително изместване. Тук статистическата теория не може да ни помогне. Възможни са само позоваванията, при това не напълно убедителни, на миналия опит" (пак там).

Аз обаче смятам, че статистическата теория може да помогне доста, но не „класическата" статистическа теория (която наистина не може да помогне), а една друга парадигма в статистиката, позната като бейсовска статистика.

Тук няма да навлизам в дълбочина в разликите между двете парадигми[3], само ще отбележа, че различията се отнасят единствено до начина на правене на изводи, базирани на извадки. Важно последствие от тези различия обаче е, че бейсовската статистика позволява да се изчисляват грешки и при непредставителни извадки.

А в конкретния случай реализираната извадка може да се разглежда като непредставителна извадка от планираната, получена по метода на отзовалите се. Тогава основен става въпросът доколко реализираната извадка достатъчно добре възпроизвежда планираната извадка. Или иначе казано, каква е грешката, но не грешката на реализираната извадка спрямо цялата генерална съвкупност, а грешката на реализираната спрямо планираната извадка. Основното допускане е, че планираната извадка е формирана по всички правила за правене на представителни извадки. Тогава, ако относителните дялове в реализираната извадка се различават малко (т.е. грешката е малка) спрямо планираната, то реализираната извадка ще е достатъчно добра.

В предишна моя публикация съм показал как се правят изводи за относителни дялове, базирани на непредставителни извадки, в два случая – на малка генерална съвкупност (Харалампиев 2004: 54-58) и на голяма генерална съвкупност (Харалампиев 2004: 66-68 и 71-79). Особеното на втория случай е, че имплицитно се допуска, че делът на извадката спрямо генералната съвкупност е пренебрежимо малък. Това допускане е напълно оправдано при всички социологически изследвания, тъй като обемите на извадката обикновено са от порядъка на (няколко) хиляди, а обемите на генералната съвкупност – от порядъка на (няколко) милиона, така че делът на извадката наистина е пренебрежимо малък.

[3] Основните разлики съм описал в предишна моя публикация (Харалампиев 2008: 146-153).

За конкретния проблем, който се разисква в настоящата статия, очевидно нито първия, нито втория случай е подходящ, защото делът на реализираната спрямо планираната извадка не е пренебрежимо малък. Затова трябва да се разработи трети случай – на голяма генерална съвкупност, при която делът на извадката не е пренебрежимо малък.

Ако означим с n обема на планираната извадка, а с $\hat{n}$ обема на реализираната извадка, то делът на реализираната спрямо планираната извадка (т.нар. response rate) ще бъде $\frac{\hat{n}}{n}$, а $1-\frac{\hat{n}}{n}$ ще е делът на отказите.

Ако се модифицират съответните формули (Харалампиев 2004: 54-55), се получава следното:

Минималната възможна стойност на относителния дял в планираната извадка е:

(1) $$p_{\min} = \hat{p}.\frac{\hat{n}}{n},$$

където:

p е относителният дял на интересуващото ни значение на изследвания признак в планираната извадка;

$\hat{p}$ е относителния дял на интересуващото ни значение на изследвания признак в реализираната извадка.

Максималната възможна стойност на относителния дял в планираната извадка е:

(2) $$p_{\max} = \hat{p}.\frac{\hat{n}}{n} + 1 - \frac{\hat{n}}{n}$$

Още тук можем да направим първия извод: целият диапазон на възможните стойности на относителния дял в планираната извадка е:

(3) $$p_{\max} - p_{\min} = 1 - \frac{\hat{n}}{n}$$

Тоест делът на отказите съвпада с диапазона на възможните стойности. Иначе казано, ако делът на отказите е 5%, тогава диапазонът на възможните стойности ще бъде 5 процентни пункта, а ако делът на отказите е 50%, тогава диапазонът на възможните стойности ще бъде 50 процентни пункта.

Този извод обаче е твърде консервативен. Това е така, защото различните стойности в диапазона не са еднакво вероятни. Затова може да се построи

доверителният интервал на относителния дял в планираната извадка и да се търси връзка между неговата ширина и делът на отказите.

Първата стъпка за построяването на доверителен интервал е да се определи т.нар. функция на разпределение. Затова продължаваме с модифицирането на формулите от цитираната публикация (Харалампиев 2004: 55):

$$(4) \qquad P(p \le x) = F(x) = \lim_{n\to\infty}\left(1 - \frac{C^{m-1}_{n-\hat{n}+m-nx+\hat{n}\hat{p}-1}}{C^{m-1}_{n-\hat{n}+m-1}}\right) = 1 - \left[\frac{1 - x - \frac{\hat{n}}{n}(1-\hat{p})}{1-\frac{\hat{n}}{n}}\right]^{m-1},$$

където m е броя на значенията на изследвания признак.

Тъй като функцията на разпределение всъщност е вероятността интересуващия ни относителен дял в планираната извадка да не надхвърли дадена стойност, то доверителния интервал се получава като разлика от две функции на разпределение:

$$(5) \qquad P(x < p \le y) = F(y) - F(x) = 1 - \alpha,$$

където α е рискът за грешка.

В конкретния случай обаче формула (5) не е приложима в този си вид. Това е така, защото формула (4) е функция на разпределение на т.нар. L-разпределение. Това е разпределение, чиято най-вероятна стойност е минималната възможна стойност и всяка следваща стойност е по-малко вероятна от предходната. В този случай е по-подходящо доверителният интервал да се затвори само отгоре, т.е.:

$$(6) \qquad P(p \le x) = F(x) = 1 - \alpha$$

Ако функцията на разпределение от формула (4) се замести във формула (6) и полученото уравнение се реши спрямо x, се получава:

$$(7) \qquad x = \left(1 - \frac{\hat{n}}{n}\right)\left(1 - \sqrt[m-1]{\alpha}\right) + \hat{p}.\frac{\hat{n}}{n}$$

Тогава ширината на доверителния интервал ще бъде[4]:

[4] При представителни извадки разпределението на относителния дял в извадката е нормално и доверителният интервал се получава като към оценката на относителния дял се прибавя и се изважда максималната грешка (Δ). Затова ширината на доверителния интервал е 2Δ. В нашия случай разпределението не е нормално и доверителния интервал не се определя като оценката плюс/минус максималната грешка, но означението 2Δ е запазено за удобство.

(8) $$2\Delta = x - p_{\min} = \left(1 - \frac{\hat{n}}{n}\right)\left(1 - \sqrt[m-1]{\alpha}\right)$$

Тази формула е достатъчна за изчисляването на ширината на доверителния интервал при всяко конкретно изследване, и на тази основа, за оценка на това с колко относителният дял в реализираната извадка се различава от планираната.

Обаче формула (8) може да се пререши спрямо дела на отказите:

(9) $$1 - \frac{\hat{n}}{n} = \frac{2\Delta}{1 - \sqrt[m-1]{\alpha}}$$

Формула (9) е полезна, защото тя дава теоретичен отговор на въпроса колко е допустимия размер на дела на отказите при предварително зададени от нас изисквания за ширината на доверителния интервал и за риска за грешка.

Тъй като в практиката на социологическите изследвания има общоприети стойности за размера на грешката и за риска за грешка, може да се направи една таблица, в която за различните стойности на m са изчислени допустимите дялове на отказите. Общоприетата стойност за размера на грешката е 3 процентни пункта[5], а за риска за грешка е 5%. Получените резултати са представени в таблица 1.

Таблица 1: *Допустими размери на дела на отказите при различен брой на значенията на изследвания признак*

Брой на значенията на изследвания признак	Допустим размер на дела на отказите
2	6,3
3	7,7
4	9,5
5	11,4
6	13,3
7	15,3
8	17,2
9	19,2
10	21,2

Таблица 1 е направена до $m = 10$, защото рядко при социологически изследвания броя на значенията на изследваните признаци надхвърля 10. Но ако

[5] Тук напълно се солидаризирам със становището на Ивайло Парчев, че тази стойност се е превърнала в „магическа" за практиката на социологическите изследвания (Парчев 1998: 121). Пак там е дадено много добро обяснение за това откъде е дошла и защо се е утвърдила тази „магическа" стойност.

все пак това се случи, тогава заинтересованият читател може сам да изчисли допустимия размер на отказите по формула (9) или ширината на доверителния интервал по формула (8).

Във връзка с таблица 1 трябва да се акцентира на три важни неща:

Първо, ако в консервативната формула (3) заложим отново, че желаем диапазонът на възможните стойности да бъде равен на две максимални грешки, тогава допустимият дял на отказите ще се получи равен на 6,0%. Тоест, когато работим с доверителни интервали, а не с диапазона на възможните стойности, получаваме по-либерални оценки за допустимия дял на отказите.

Второ, обикновено в анкетната карта на всяко социологическо изследване има множество въпроси, всеки с различен брой отговори. Затова, когато определяме допустимия дял на отказите, трябва да използваме признака с най-малко на брой значения. И също така обикновено в анкетните карти се срещат дихотомни въпроси. А както се вижда от таблица 1, допустимият дял на отказите при дихотомните признаци съвсем малко се отличава от допустимия дял на отказите, получен на база на диапазона на възможните стойности.

Трето, числата в таблица 1 все пак трябва да се разглеждат повече като илюстрация, а не като твърд еталон. Това е така, защото те са изчислени при зададено изискване ширината на доверителния интервал да бъде шест процентни пункта. Но при конкретно изследване тази ширина може да се окаже или по-висока, или по-ниска от желаната. А това означава и, че праговете трябва съответно да се намалят или да се увеличат. Така например, при едно електорално изследване, за големите партии, които събират 20-30% подкрепа, доверителен интервал с ширина от шест процентни пункта може да е подходящ, но за малките партии, които събират 3-4% подкрепа, този доверителен интервал е напълно неприложим. За малките партии може би по подходящ е доверителен интервал с ширина от един процентен пункт. Но намаляването на ширината на доверителния интервал шест пъти автоматично намалява и допустимия дял на отказите също шест пъти!

При всяко социологическо изследване винаги ще има откази. Но резултатите, получени досега, дават възможност на изследователя да направи две неща. Първо, преди започване на теренното изследване може да изчисли по формула (9) допустимия дял на отказите и по време на теренната работа да удържа отказите в тези граници, и второ, след приключването на терена да

изчисли по формула (8) диапазона, в който (най-вероятно) попада разликата в интересуващите ни относителни дялове между реализираната и планираната извадка.

ЛИТЕРАТУРА

Атанасов, А. и кол. 2006. *Електоралните изследвания. Изследователски проблеми и прогностични възможности*. Издателство „FABER", Велико Търново.

Венедиков, Й. 1992. *Статистика, социология и още нещо...* Издателство „Информационно обслужване", София.

Венедиков, Й. 1993. *Общественото мнение. Епистемологични проблеми*. Университетско издателство „Св. Климент Охридски", София

Дерменджиева, Б., В. Цветкова, Н. Милева. 2010. „Отказите от участия в изследвания: проблем на обществото?". В: *Благополучие и доверие: България в Европа?*, състав. Н. Тилкиджиев и Л. Димова. Издателство „Изток-Запад", София.

Парчев, И. 1998. *Избор на партия, избор на президент. Осем етюда върху една таблица*. Статистическо издателство и печатница при НСИ, София.

Харалампиев, К. 2004. *Нетрадиционен поглед върху традиционни статистически проблеми*. Издателство „Балкани", София.

Харалампиев, К. 2007. Агенциите рядко грешат умишлено. *Вестник „Кеш", 39: 16*

Харалампиев, К. 2008. „За парадигмите в статистиката – бейсовска статистика". В: *Актуални проблеми на статистическата теория и практика*. Университетско издателство „Стопанство", София.

Биографична справка: Калоян Харалампиев е доктор по икономика, специалност „Статистика и демография". Доцент в катедра „Социология" на Философски факултет на СУ „Св. Климент Охридски", преподавател по „Статистически методи в социологията" и SPSS. Дисертация „Трудов потенциал на населението на България – минало, настояще и бъдеще" (2003). Научни интереси основно в областта на теорията на вероятностите, бейсовската статистика и демографската статистика.

Основни публикации: книгите *Въведение в основните статистически методи за анализ* (учебник) (2003), *Трудов потенциал на населението на България (1992-2001)* (2003), *Нетрадиционен поглед върху традиционни статистически проблеми* (2004), *SPSS за напреднали* (учебно помагало) (2007) и *Работа с данни с SPSS* (учебно помагало) (2009).

Адрес за кореспонденция:
София 1113, Цариградско шосе №125, бл. 4, каб. 413
Сл. тел.: 971-10-02 (вътрешен 320)
Моб. тел.: 0885-04-62-56
E-mail: k_haralampiev@hotmail.com, kaloyan_haralampiev@yahoo.com