КАЛИБРАЦИЯ НА ДАННИ ОТ СОЦИАЛНИ ИЗСЛЕДВАНИЯ

Йордан Калчев, Вера Велева***

Въведение

Извадковият подход намира широко приложение при изследване на социалните процеси. Независимо че теорията на извадковите изследвания е много добре разработена, при тяхното практическо осъществяване възникват редица проблеми. Една част от тях са свързани с необхващането по различни причини на всички единици от планирания обем на извадките. Този необхват често води до нарушаване на представителността и точността на получаваните данни от извадките, които се използват при провеждане на емпиричните изследвания.

Целта на изследването е да представи възможност чрез съвременните средства на статистиката и математическото моделиране тези недостатъци в извадковите изследванията да бъдат преодолени. Такива средства са моделите за калибриране на данните. Те са нов подход в обработката на данните от извадковите изследвания и имат своята научна обоснованост за неутрализиране на влиянието на различни смущаващи фактори върху изследователския процес.

Крайната цел на всяко емпирично изследване е да възпроизведе в максимално приближена степен изследвания процес, да установи и измери неизвестни негови страни и параметри. Това налага широкото приложение на различни статистико-математически методи по време на целия процес на провеждане на социалните изследвания. Основа за използването на методите на статистиката в социалните изследвания е съществуващата обективност, че обектите на социалните изследвания се проявяват като специфични статистически структури. „Спецификата на обществените явления не е пречка за изучаването им чрез случайни извадки, тъй като повечето обществени явления не зависят от случайни фактори и са обект на съзнателна човешка дейност. Дори когато генералната съвкупност е била напълно обусловена от човешка намеса и нито един случай на съвкупността не е настъпил случайно, не съществува пречка изучаването на сводните свойства на съвкупността да стане

* Д-р, доц., катедра „Социология”, ЮЗУ „Неофит Рилски”; e-mail: ikaltchev@abv.bg.

** Главен асистент, катедра „Социология”, ЮЗУ „Неофит Рилски”; e-mail: veleva_v@abv.bg.

непреднамерено, т.е. несъзнателно избиране на случаите на извадката“ (Цонев, 1970, с. 74).

Изследванията (Калчев, 2005, с. 70) са „научно обосновани само тогава, когато получената информация от включените в наблюдението единици, чрез формираната извадка дава възможност получените от извадката характеристики да се отнасят (важат) за цялата генерална съвкупност. Това теоретично е допустимо и възможно само когато формираната извадка е репрезентативна (представителна) на изучаваната генерална съвкупност, т.е. приложени са правилата на статистическия метод на изучаване“.

За осигуряването на представителността на данните в практиката на емпиричните изследвания се използват различни организационни техники и методи за формиране на извадките, тъй като в социалната действителност невинаги може да се приложи простият случаен подбор в чист вид. При това в някои случаи отстъплението от строгите правила на простия случаен (вероятностен) подбор дори подобрява резултатите, но това става само при определени правила. Има се предвид изпълнението на процедури, които не само оптимизират разходите за провеждане на изследването, но и постигат по-добри статистически резултати, или реализиране на т.нар. „дизайн ефект“ - използване на степенен гнездови подбор, стратификационни процедури, съчетаване на статична с динамична представителност, отчитане на рискове за реализиране на извадките текущ контрол във фазата на наблюдението.

Посочените вече условия са важни, но с това не се изчерпва задачата по формирането на представителни извадки. Трябва да се има предвид, че независимо от спазването на всички тези правила всяка извадка носи в себе си определена грешка - случайна грешка, произхождаща от това, че се изследва част от съвкупността. Тази грешка основно зависи от обема на извадката. С увеличението на извадката грешката намалява и то неравномерно. Има се предвид, че връзката между размера на грешката и обема на извадката е опосредствана чрез радикала $\sqrt{n}$. Тази връзка е важна и е в основата на доказване на достатъчна представителност при минимален обем на извадката.

Известно е, че този вид грешки при представителните извадки мо-

гат да бъдат изследователски предвидени и контролирани още в процеса на проектиране на конкретната извадка. Съществува обаче, особено в сегашната социална действителност, един голям проблем, който е свързан с нереалното изпълнение на планираните извадки. Причините могат да бъдат най-различни - липса на изследователски капацитет, лоша социална обстановка, ниско ниво на сътрудничество от страна на респондентите, намеса на външни (съзнателни и несъзнателни) фактори, възпрепятстващи изследването, недобра организация на изследователския процес и други. В резултат на това не се постига изпълнението на планирания обем единици за наблюдение и оттук е нарушена структурата на изследваната съвкупност. Тези грешки се определят като грешки, породени от непълнотата на обхвата на извадката. Те могат да застрашат изпълнението на извадковото наблюдение, признаването на неговите резултати и заключения, да поставят не само под съмнение, а и напълно да компрометират, представителността и точността на получените резултати.

Такива проблеми в определена степен винаги са съществували. Затова са използвани различни методи за оценка и претегляне на резултатите от получаваните извадки, подсилване на отделни групи в извадките, добавяния на записи, отчитане на вътрешногнездовите корелации, автоматични корекции на данни чрез величини за тенденциите и т.н. При сегашните условия на все по-трудно събиране на данни от респондентите вече са разработени и се прилагат нови математико-статистически методи и модели за редактиране и запълване на липсващи данни. Тяхното използване ще се налага не само от потребностите на традиционните представителни извадкови изучавания за събиране на данни, но и поради все по-честото използване на новите технологии в изследванията - телефонно интервю (CATI), уебизследвания по интернет и други електронни форми.

Често в практиката на социалните изследвания при анализа на резултатите се игнорира загубата на данни от извадката, а по този начин и влиянието на грешките, които възникват допълнително. Причината е изследователска некоректност или допускането, че разпределението на неучаствалите в изследването не се различава от това на участниците по изследваните характеристики. В противен случай, ако разпределенията

на неотговорилите и отговорилите са значимо различни, то резултатите ще бъдат обременени и с нестохастична систематична грешка.

1. Липсващи данни поради необхват на единиците в извадката

В теорията и практиката са разработени различни подходи за намаляване на броя на необхванатите единици преди и по време на теренната работа. Предприемат се организационно-технически дейности, които целят осигуряването на изпълнима методика на извадката. Разбира се, те дават резултати, но невинаги могат да компенсират загубите на информация, а освен това изискват допълнителни ресурси и време.

Проблемът за нестохастичните грешки е твърде важен, тъй като присъства при всяко изследване. В това изследване се ограничаваме само до една част от тях - грешките, свързани със загубата на данни при събиране на първичната информация, т.е. липсата на данни за отпадналите от извадката единици. Основният въпрос тук е кога и при какви условия изпълнението на извадката може да влоши качеството на получаваните резултати от конкретното изследване до такава степен, че резултатите да загубят своята представителност и да нарушат очакваната точност.

Този въпрос не може да получи еднозначен и стандартизиран отговор. Затова е необходимо за всяко конкретно изследване да се подложат на преценка някои преки резултати от проведеното наблюдение. Необходимо е да се установят броят и относителният дял на необхванатите (ненаблюдаваните) единици от извадката. Трябва да се допусне, че съществува вероятност за съществени различия в информацията, която предоставят участниците в изследването и тези, които не участват в него. Тези различия не може да се очаква, че са резултат на действието само на случайни фактори. Освен това, трябва да се има предвид, че свиването на обема на извадката води и до нарастване на стохастичната грешка, което също не е проблем за подценяване.

2. Влияние на необхвата върху точността на оценките

Предвиждането и оценяването на стохастичната грешка не е проблем за подготвения изследовател. Оценяването на размерите на нестохастичната грешка в извадковите изследвания обаче е твърде сложен и труден проблем.

Теоретично се знае, че средната аритметична величина е неизместена оценка. При необходимост величината на нейното изместване, по-

лучено от проста случайна извадка, може да се установи по формулата:

$\bar{Y}_1 - \bar{Y} = \bar{Y}_1 - (W_1.\bar{Y}_1 + W_2.\ \bar{Y}_2) = W_2(\bar{Y}_1 - \bar{Y}_2),$

където:

$\bar{Y}_1$ - средна, изчислена за участвалите в извадката;

$\bar{Y}_2$ - средна, изчислена за неучаствалите в извадката;

$\bar{Y}$ - средна за цялата съвкупност;

W_2 - относителен дял на необхванатите в извадката. Знае се също, че: $W_2 = 1-W_1$; W_1 - относителен дял на обхванатите в извадката (Кокрен, 1976, с. 381).

От формулата може да се резюмира, че изместеността на средната оценка в резултат на необхвата е произведение между относителния дял на неучастващите в извадката и разликата между средните на участвалите и неучаствалите в извадката. Тъй като $\bar{Y}_2$ е неизвестно, няма информация за границите на тази променлива и не може да се пресметне значението на изместването. Това може да се постигне, ако е налична информация за тази променлива от други източници - изчерпателни изследвания, регистри, стандарти, които могат да се приемат за точни.

В практиката на емпиричните социални изследвания най-често се оценяват относителни дялове. Известно е равенството $p+q=1$. Оттук границите на q вече са известни и са между 0 и 1. Тогава границите на оценявания относителен дял (P) ще бъде:

$$P = \pm Z * \sqrt{\frac{pq}{n_1}} * \sqrt{1 - \frac{n}{N}},$$

където:

P - оценяван относителен дял в извадката;

Z - гаранционен множител;

P - оценката на P от извадката;

$q = 1 - p$

n - планиран обем на извадката;

n_1 - брой на обхванатите в извадката единици: $n_1 = nW_1$;

N - брой на единиците в генералната съвкупност.

Въз основа на посочената формула Кокрен (1976, с. 381) предлага долната и горната граница на доверителния интервал да се изчисляват съответно:

$$P_L = W_1\left(p_1 - 2\sqrt{\frac{p_1 q_1}{n_1}}\right) + W_2(0),$$

$$P_U = W_1\left(p_1 - 2\sqrt{\frac{p_1 q_1}{n_1}}\right) + W_2(1).$$

От приведените формули се установява, че при нарастването на W_2 ширините на доверителния интервал за оценявания параметър P нарастват и получените резултати от извадката са твърде нестабилни. Направените по горните формули изчисления показват, че ако от извадка с обем n = 1000 единици е получена извадкова оценка p_1=*0.05,* (респ. 5%) доверителните интервали на оценявания параметър P от генералната съвкупност ще зависят от относителния дял на ненаблюдаваните - W_2.

В случаите, когато $W_2 = 0\%$, стойността на оценявания параметър ще бъде в границите 3.6 - 6.4%. Ако $W_2 = 5\%$, тези граници ще бъдат съответно 3.4 и 11.1%. С нарастването на относителния дял на необхванатите лица от извадката точността на получаваната информация намалява - увеличава се максималната грешка, респ. разширява се доверителният интервал, който може да доведе до неизползваемост на получените оценки. Например ако относителният дял на необхванатите от извадката е 15%, то границите на доверителния интервал за оценявания 5% относителен дял ще бъдат между 3 и 20.5%. При този случай максималната относителна грешка (Δ%) ще бъде 60% и 410%. Оттук следва изводът, че за да се получат оценки в задоволителни граници в определена степен може да се компенсира с увеличаване на обема на извадката (напрягане на извадката), но с това не се решава проблемът изцяло.

3. Претегляне на данни от извадката. Необходимост от претегляне

Претеглянето на данните, получени от извадката, е известна и често прилагана процедура за постигане на критериите за качество на информацията. Тя е особено необходима, когато резултатите от проведеното емпирично изследване се разпростират (обобщават) за цялата генерална съвкупност. Претеглянето на данните се използва, за да се намали систематичното отклонение при оценките от извадковото изследване (нестохастичната грешка). В него е залегнала хипотезата, че вероятностите за получаване на отговор се различават при двете части на извадката - на участвалите и неучаствалите в извадката лица.

Претеглянето се извършва поради необходимостта да се игнорират преднамерените и намалят непреднамерените отклонения от планираната равна вероятност за попадане на единиците от генералната съвкупност в извадката. Например ако дадена страта (териториален район, социална група) е свръхпредставена в извадковото изследване, е възможно като се използват допълнителни данни (например данни от официалната статистика, официални регистри) да се изравни извадковото разпределение с разпределението на параметрите в генералната съвкупност.

В статистиката се предлагат различни и известни решения за претегляне на данни от извадки, обединени под наименованието „методи за претегляне на извадкови данни". Става дума за процедури, чрез които се цели да се постигне приближение или изравняване на извадковото разпределение до разпределението на генералната съвкупност.

Най-общо процедурата по претеглянето на данните включва изчисляване на тегла (W_i) обратно пропорционални на вероятността за включване на всяка единица в извадката (P_i) - (базови тегла): $W_i = \frac{1}{p_i}$.

Тези тегла позволяват да се: 1) отчетат параметрите на извадката и да се компенсират различията; 2) да се направи корекция на неотговорилите в извадката (преразпределяне на теглата на неотговорилите сред респондентите); 3) обвързване на оценките с известни параметри на генералната съвкупност (Kish, 1965).

4. Корекция на данните чрез методите на постстратификацията

Конвенционалният подход за претегляне на данните от извадки предполага, че за всички единици и за всяка от тях са събрани пълни данни в масива. На практика такава идеална ситуация трудно може да бъде реализирана. Затова се предлага и прилага друг подход за решаване на проблемите във връзка с неполучаване на отговори от изследваните единици - групово претегляне (weighting classes).

Идеята за приложението на постстратификацията се основава на наличието на информация за основни демографски и социални признаци (пол, възраст, местоживеене, образование, икономическа активност, административни единици) за цялото население, въз основа на които се формират различни страти сред изследваната съвкупност.

Аналитично същността на постстратификацията най-общо може

да се представи чрез формулата:

$$K_m = \frac{N_m}{\sum W(h_i)j * \delta(j)},$$

където:

K_m - постстратификационен коефициент за m-ата страта;

N_m - брой на единиците в m-ата страта сред цялото население;

$W(h_i)j$ - тегло с поправка на неотговорилата единица j;

$\delta(j)$ - множител със стойност 1 (едно), когато единицата принадлежи към дадената страта и 0 (нула), ако не принадлежи[1].

Сумата от теглата в знаменателя обхваща всички единици, които са дали информация в извадката. Тегло за поправката на неотговорилите от всички отговорили от стратата m се умножава с Km за получаване на постстратификационното тегло.

При използването на такъв подход трябва да се има предвид, че в отделните групи (страти) трябва да има достатъчен брой единици (напр. > 30), осигуряващи стабилност на изчисляваните показатели. Също трябва да се има предвид размерът на поправката, която ще се извършва за всяка страта, изборът на променливата, която формира стратите, и други.

В изследователската практика за постстратификацията съществуват многовариантни продължения. Такъв вариант е наличието на контролна информация за изследваната съвкупност по два и повече признака. Подходът в такъв случай изисква изравняване на теглата последователно по тези измерения. След тази процедура сумата от корекциите на теглата в максимална степен ще съответства на използваните контролни показатели - за всеки признак поотделно, но може да не съответства на комбинацията от двата признака.

Когато са налични няколко контролни променливи, могат да се използват и по-сложни методи, като се изследват линейни зависимости напр. регресионен анализ. В този смисъл постстратификацията може да се разглежда като един „несложен праволинеен метод за калибрация на

[1] Руководство по подготовке статистических данных об использовании времени для оценки оплачиваемого и неоплачиваемого труда (ST/ESA/STAT/SER.F/93), Департамент по экономическим и социальным вопросам, Статистический отдел ООН, Нью-Йорк, 2007 год., с. 159 - 160.

данни”[2] от извадковите изследвания.

5. Калибрация на данни от извадкови изследвания. Същност

Идеята за калибрацията на данни от извадковите изследвания получава голям напредък след публикуването в края на миналия век на поредица от статии на специалисти (Deville, Sarndal, 1992; Ekholm, Laaksonen, 1991; Lundstrom, Sarndal, 1999) в областта на теорията и практиката на статистическите извадкови изучавания. Според посочените автори калибрацията осигурява системен възглед за изследванията дори и в присъствието на различни грешки, несвързани с проектирането на извадката.

В теоретичната основа на груповото и постстратификационното претегляне стоят инверсиите на вероятностите за включването на единиците в извадката. Калибрацията също използва тези идеи, но ги специфицира и доразвива.

Калибрацията (Calibration) като метод за селектиране на тегла в извадковите изследвания (извадкови тежести) е по-разпространена в статистическите служби на редица страни. В последно време Евростат все повече препоръчва тази практика да бъде използвана от държавите - членки на ЕС, при обработката на данни от извадковите изследвания.

Направеното проучване на различни литературни източници във връзка с процедурата за калибрация на данните показа, че съществува голям брой публикации по тази тематика. Значителна част от тази литература е трудна за възприемане, тъй като в нея на много високо теоретично и аналитично (математическо) ниво се разглежда същността на калибрацията на данни. Поради това ние ще се спрем на нейната същност предимно от позицията на нейните функции и практическо приложение при решаване на проблемите за точността на данните от извадковите изследвания. Това ограничение се оправдава от по-голямата важност на практическите ефекти от прилагането на калибрацията и основно от факта, че всички процедури за калибриране на данни са налични и разпространени в софтуерните продукти - SAS, SPSS и други.

Calr-Erik Sarndal, разглеждайки различни варианти за калибриране на данни в статистическите изследвания (Deville, Sarndal, 1992; Sarndal, 2007), определя метода като „подход за оценка на крайната съвкупност

[2] Пак там, с. 160.

и се състои от:

а) изчисляване на тегла (тежести), които включват определена помощна информация и помощни уравнения за калибриране;

б) използването на тези тегла, за да се изчислят линейно претеглените оценки на суми и други крайни параметри на съвкупността;

в) получаването на обективни, почти безпристрастно проектирани оценки, дори и когато има „nonresponse" и други грешки от извадката".

Най-общо в литературата „калибрацията" се определя като метод за повторно претегляне и се използва, когато има достъп до няколко променливи, качествени и количествени, на които трябва да се извършва съвместна корекция. Калибрирането се представя като набор от тегла за единиците в извадката, които отговарят за калибриране на известни крайни обобщаващи за съвкупността величини и такива, които оценителят (теглата) довежда до съответствие (т.нар. проектиране в съответствие) със съвкупността, или че дизайнът има асимптотично незначителен принос към средната квадратна грешка на оценителя (Kott, 2006). С други думи, целта е постигане на безпристрастно проектиране и изпълнение на извадката. „Калибрацията чрез известни (налични) демографски показатели се използва за намаляване на изместеността на оценките, свързана е с обхвата и не може да се отстрани чрез претегляне и съгласуване на оценките с известни показатели. Тя позволява да се съкрати грешката на извадката в оценяването, като добре се корелира с налични контролни показатели"[3].

В разпространените от Статистическия институт на Канада (The Quality Guidelines..., 2003) Указания за качество се посочва, че калибрацията е процедура, която може да се използва за включване на допълнителни данни за регулиране на теглата чрез използване на множители, известни като калибровъчни коефициенти, които правят оценките съгласувани с определени обобщаващи величини на изследваната съвкупност. Получените тегла се наричат „калибриращи тегла" или „окончателни тегла за оценяване".

Калибрирането на данните се разглежда като нов, системен подход, който взима под внимание спомагателна информация за съвкупността и в стандартни условия осигурява включването на спомагателната ин-

[3] Пак там, с. 160.

формация в оценката. „Калибрацията (Степанов, 2009, с. 1) е процес на целенасочено изменение на параметрите на извадковия план както на извадковите тегла, така също и на непараметричните модификации в състава на извадката за намаляване на извадковите грешки и повишаване на точността и устойчивостта на извадковите оценки на статистическите показатели", т.е. калибрирането е специфичен начин да се отчете налична за изследваната съвкупност помощна информация и чрез различни стандартни настройки тази информация да бъде включена в получаването на крайните оценки. Същността на извършваното претегляне чрез процедурите за калибрация се състои в това, че теглата за всяка наблюдавана единица се изчисляват и присвояват по всички калибриращи признаци едновременно.

Могат да се посочат още редица определения и ефекти от калибрацията на извадковите данни, в които не се открояват съществени различия. Ако има такива, те се отнасят до това, че в някои от посочените и съществуващите в литературата определения се акцентира върху теоретичната база, а в други - на процедурния характер на калибрацията.

Приведените дефиниции за калибрацията относно нейната същност показват, че тя осигурява системно разглеждане на изследванията относно параметрите на извадката и изследваната съвкупност, „даже в присъствието на различни грешки, несвързани с планираната извадка" (Степанов, 2009, с. 6).

Въз основа на посочените определения за същността на калибрацията може да се заключи, че калибрацията представлява статистическа процедура, чрез която се преодоляват проблемите, свързани с липсващите данни за някоя променлива и проблемите, породени от неизпълнение или свръхнабиране на единици за отделни групи от съвкупността. В резултат на балансиране на цялата извадкова съвкупност в съответствие с нейни обобщаващи характеристики („съгласувани оценки") се осигурява непреднамереността на подбора и се намалява изместеността на получаваните оценки. Така в крайна сметка се постига по-голяма точност в стойностите на обобщаващите параметри на генералната съвкупност.

За процеса на калибрация се използват както наличните данни от дадено извадково изследване, така и допълнителна информация за характеристики на генералната съвкупност, като комбинирането им се извършва посредством калибрационни уравнения (функции). Използва-

нето на такава процедура е възможно при наличието на информация от допълнителни източници, която да характеризира в различни аспекти изследваната съвкупност - това са различните административни регистри, статистически преброявания или други изчерпателни записи за единиците в съвкупността.

6. Аналитичен модел за калибрация на данни

Общо универсално аналитично представяне на разработените модели за калибриране на данните е трудно да се направи, тъй като те са сложни математически конструкции и съществуват различни варианти на моделите. Един опростен вариант на модел за калибрация на данни може да се представи в следния вид[4]:

Приема се, че обект на изследване е наличната генерална съвкупност U, която се състои от N елементи, като $U = \{1,..., k, ..., N\}$. От проведеното изследване е необходимо да се оцени значението на променливата J. Нека с $y_k = (y_{k1}, \dots, y_{kJ})'$ означим вектор, значенията на променливата Jх 1, за k-я елемент на съвкупността.

В изследването крайната цел е получаването на оценка на вектора от всички сумарни значения за променливата:

$$\mathbf{T}_y = \sum_{k \in \mathbf{U}} \mathbf{y}_k = \mathbf{Y}'_{\mathbf{U}} \mathbf{1}_N,$$

където:

с Y'_U е означена матрица с Nх1 значения на y.

При формирането на извадката се допуска, че n единици от съвкупността U са включени в извадката s. За всяко извадково изследване крайната цел е да се получат оценки T_y. Ролята на „стандартна" оценяваща формула на сумарните данни се изпълнява от оценъчната формула на

[4] За представяне на модела са използвани публикациите:
Task Force on the implementation of NACE, Rev. 2 - Handbook on methodological aspects related to sampling designs and weights estimations, Eurostat, 2006.
S. Lundstrom (Statistics Sweden), C. Sarndal (Statistics Canada), Calibration as a Method for Deriving Nonresponse Adjusted Weights.
C. E. Sarndal, The calibration approach in survey theory and practice, Statistics Canada, Vol. 33, No 2, 2007, p. 99 - 119.
С. В. Степанов, Калибровка выборки, М., 2009.

Хорвиц - Томсън (H-T), която се определя като:

$$\hat{\mathbf{T}}_y = \sum_{k \in s} d_k \mathbf{y}_k,$$

където:

$d_k = \frac{1}{\delta_k}$ е извадково тегло за единицата k, а δ_k е вероятност за включване на единицата k в извадката[5].

Въпреки равните възможности за включване на всички единици от съвкупността U в извадката s по различни причини липсват отговори (nonresponse). Сред различните способи за намаляване на влиянието на неучастието на избраните единици в изследването върху получаваните оценки калибрацията на данните дава добри възможности.

Много често при проектирането на извадкови изследвания са налични променливи (x_k), които са обвързани с променливата y. Тази обвързаност и наличните данни за тези променливи се използват за подобряване на точността на оценките на изследвания параметър - T_y.

Цялата тази допълнителна информация може да се означи и представи като вектор X_k. За да продължи процедурата, е необходимо да се състави т.нар. оценител[6].

Същността на процеса на калибриране на данните накратко се изразява в следното. Приема се, че са известни сумарните (обобщаващите) оценки на съвкупността за променливите x и че те трябва да се оценят по данни от извадката, като се използва оценяващата формула (H-T). Това предполага да се получат оценки T_x чрез:

$$\hat{\mathbf{T}}_x = \sum_{k \in s} d_k \mathbf{x}_k.$$

Получените оценки $\hat{T}_x$ обаче невинаги могат да осигурят точково съвпадение (точково оценяване) със съответните сумарни данни T_x на съвкупността, което се изразява в отклонението $\hat{T}_x - T_x$.

За да се избегне тази разлика (грешка на калибрацията), оценяващата формула може да се промени, като се заменят извадковите тегла d_k с нови тегла w_k, които се включват в калибровъчната формула:

[5] Task Force on the implementation of NACE, Rev. 2 ..., Eurostat, 2006, Annex A, p. 16.
[6] За по-подробно запознаване с функциите и видовете оценители вж.: S. Lundstrom (Statistics Sweden), C. Sarndal (Statistics Canada), Calibration as a Method for Deriving Nonresponse Adjusted Weights (Statistics Sweden).

$$\hat{\mathbf{T}}_{xC} = \sum_{k \in s} w_k \mathbf{x}_k,$$

където:

(w_k, x_k) са изследваните тегла без отклонения в калибровката и удовлетворяват равенството:

$$\hat{\mathbf{T}}_{xC} - \mathbf{T}_x = \sum_{k \in s} w_k \mathbf{x}_k - \mathbf{T}_x = \mathbf{0}.$$

Смисълът на използването на този оценител е, ако използваните калибровъчни тегла (w_k) са в състояние да намалят или ако е възможно да отстранят грешката при получаване на сумарните оценки на *x,* то те могат да се използват в оценъчната формула при калибрацията и да намалят грешката при получаването на сумарните оценки за *y*:

$$\hat{\mathbf{T}}_{yC} = \sum_{k \in s} w_k \mathbf{y}_k.$$

Това аналитично представяне на калибрацията на данни показва само общата идея на тази методология. Съществуват различни модели (оценяващи формули), които се прилагат в статистическата практика.

Най-често използваните калибрационни функции са:

- линейна функция: *f(x) = y= a+bx;*
- експоненциална функция: $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$;
- прекъсната (ограничена, отрязана*) линейна функция;

*(*Функцията f(x) е ограничена* ***в*** интервала *(a, b), ако съществуват две числа A, B* такива, че $A \leq f(x) \leq B$ за всяко $x \in (a, b)$.)

- логит функция (ограничена): $logit(p) = log\left(\frac{p}{1-p}\right)$.

Независимо от тяхното разнообразие и сложен математически апарат те са достъпни за по-широк кръг специализирани потребители, тъй като имат своите софтуерни аналози в готови програмни продукти.

Известни калибрационни програми са: CALMAR, GES, BASCULA, CLAN и g-CALIB[7]. От посочените пет програми беше установено, че по-достъпна за приложение е програмата g-CALIB, която е базирана в софтуерния продукт SPSS. Останалите основно са базирани в SAS софтуер, а той не е широко разпространен в нашата страна.

7. Калибрация на данни от проведено емпирично изследване

За пример ще използваме „Епидемиологично изследване на насилие на деца в балканските страни" (BECAN), което е реализирано от

[7] Цялата математическа разработка и програмно осигуряване на тези подходи в сравнителен аспект могат да се видят в: C. Vanderhoeft, Generalised Calibration under SPSS, g-CALIB, Statistics Belgium, 2002.

екип в ЮЗУ „Неофит Рилски".

За реализацията на поставените основни цели и аналитични задачи по този проект е проведено извадково изследване. Обект на изследване са ученици, обхванати във формалното образование, на възраст 11, 13 и 16 години. Изследването е реализирано чрез извадка от определените целеви съвкупности, обект на изследването. Наблюдението е проведено в три области на страната.

Използваната извадка е излъчена пропорционално на броя на изследваната съвкупност във всяка област. За всяка област извадката е стратифицирана по местонамиране на училищата (град, село) и по възраст на изследваните лица. За подходящ модел на извадката е възприет този на двустепенната гнездова извадка - с гнездо на първа степен училището и на втора степен паралелките (класове), като в тях са анкетирани всички ученици. Целта е избраният модел да осигури представителност и необходимата точност на резултатите общо, по области и възрасти.

В представянето на този пример за калибрация на данните се правят сравнения на резултатите, получени: пряко от изследването, от просто претегляне - с тегла от генералната съвкупност, и резултати след изпълнение на процедури за калибрация на данните.

Изпълнението на процедурите по претеглянето и калибрацията на данните от това извадково изследване включват:

1. Определяне на признаците, за които има данни за цялата генерална съвкупност и върху които могат да се приложат процедури по претегляне и калибрация на резултатите. За претеглянето на резултатите от изследването и изчисляването на калибрационните тегла са използвани статистически данни[8] за броя и разпределението на учениците по възраст в областите и по местонамиране - град, село. Общият брой (генерална съвкупност) на учениците в трите области от посочените възрасти за изследвания период (2009 - 2010 г.) по официални статистически данни е малко над 28 хиляди.

2. Изчисляване на относителните дялове по разновидностите на тези характеристики за общата съвкупност и за извадковата съвкупност.

3. Сравняване на тези характеристики и вземането на решение за претегляне и калибриране на събраните данни. Резултатите от направе-

[8] По данни на НСИ към октомври 2009 година.

ния анализ показаха, че е необходимо да бъдат предприети такива процедури с цел по-добро възпроизвеждане на основни структури на генералната съвкупност.

4. Избор на софтуер и модел за претегляне и калибриране на данните от изследването. С разполагаемата софтуерна среда у нас, единствено приложим е продуктът g-CALIB, който е базиран в SPSS.

5. Въз основа на получените данни от извадковото изследване и наличните данни за генералната съвкупност беше подготвен информационен файл съобразно изискванията на избрания софтуер и беше включен в него.

6. Претегляне на данните от изследването и калибрация на данните (изчисляване на калибрационни тегла).

Обикновеното претегляне на данните е извършено, като е използвана наличната статистическа информация за разпределението на изследваната съвкупност по избраните признаци. Използвана е познатата процедура, която включва използването на тегла (*Wi)* обратно пропорционални на вероятността за включване на всяка единица в извадката (*Pi)*.

7. Претегляне на данните от извадката и генериране на многомерни разпределения. След провеждане на процедурите по претеглянето и калибрирането на данните, са произведени таблици, върху които е извършен анализ.

При сравняването на получените от извадката относителни дялове за възрастовите групи и местонамирането на училищата бяха установени различия (табл. 1).

1. Дял на общо анкетираните ученици в градовете в отделните възрасти

(Проценти)

Възраст	Резултати от анкетата	Статистически данни (НСИ)
11 години	72	75
13 години	69	75
16 години	73	94

Аналогични са резултатите и по отношение на учениците, които учат в селата.

Установените различия по тези стратифициращи признаци показват необходимостта от претегляне на данните (корекция на теглата, с ко-

ито участват отделните страти), преди да се пристъпи към обработката на данните и тяхното анализиране.

Резултатите от изпълнените процедури по претеглянето на данните и калибрацията са представени в табл. 2.

2. Двумерно разпределение на изследваните лица по възраст, местонамиране на училището и по области

Възраст	Области	Непретеглени данни		Тегла, коригирани с неотговорилите (non-respons)		Калибрирани тегла		Общо
		населено място		населено място		населено място		
		град	село	град	село	град	село	
а	б	1	2	3	4	5	6	7
11 г.	Благоевград	56.5	43.5	71.6	28.4	64.2	35.8	100.0
	Варна	79.9	20.1	86.2	13.8	82.0	18.0	100.0
	Велико Търново	66.2	33.8	81.6	18.4	75.5	24.5	100.0
	Общо	**71.5**	**28.5**	**81.5**	**18.5**	**74.8**	**25.2**	**100.0**
13 г.	Благоевград	56.4	43.6	71.6	28.4	62.3	37.7	100.0
	Варна	78.3	21.7	85.1	14.9	83.1	16.9	100.0
	Велико Търново	62.9	37.1	79.3	20.7	75.5	24.5	100.0
	Общо	**68.8**	**31.2**	**79.4**	**20.6**	**74.6**	**25.4**	**100.0**
16 г.	Благоевград	58.8	41.2	73.5	26.5	89.5	10.5	100.0
	Варна	86.6	13.4	91.1	8.9	96.2	3.8	100.0
	Велико Търново	72.0	28.0	85.3	14.7	95.9	4.1	100.0
	Общо	**72.7**	**27.3**	**82.8**	**17.2**	**93.9**	**6.1**	**100.0**
Общо	Благоевград	57.4	42.6	72.4	27.6	72.4	27.6	100.0
	Варна	81.2	18.8	87.2	12.8	87.2	12.8	100.0
	Велико Търново	67.5	32.5	82.4	17.6	82.4	17.6	100.0
	Общо	**71.0**	**29.0**	**81.3**	**18.7**	**81.3**	**18.7**	**100.0**

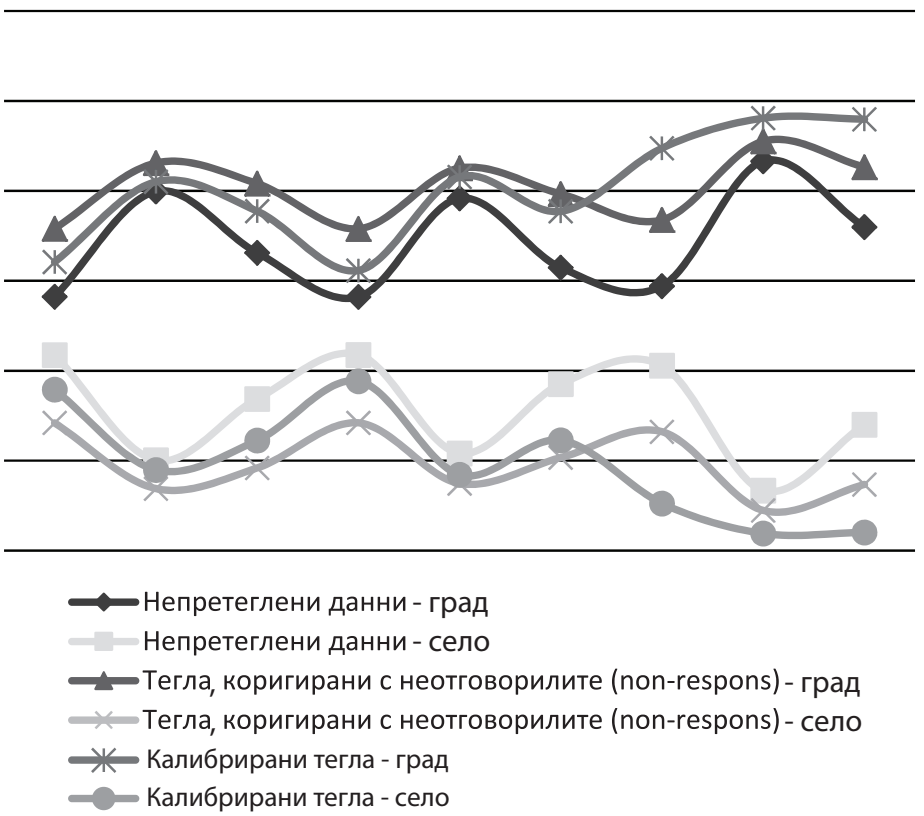
В табл. 2, в кол. 1 и 2 са посочени непретеглени данни (относителни дялове - в %), получени от изследването за всяка възрастова група поотделно за всяка област, разпределени по местонамиране на училищата - град, село. В кол. 3 и 4 е извършено обикновено претегляне, като са

използвани данни на НСИ от статистиката на образованието. От същите официални статистически данни са изчислени калибрационни тегла, които са представени в кол. 5 и 6.

При сравняването на получените относителни тегла (в %) се установява, че има различия в зависимост от начините на претегляне. Представени са общите изводи относно получените тегла и посоката на различията между тях.

Различията визуално са показани на фиг. 1.

Фиг. 1. Съотношение на теглата на изследваните страти по възраст, тип населено място и по области



От табличното и графичното представяне на получените и изчислените тегла на отделните страти се установява, че и в трите области

относителните тегла на непретеглените данни за трите възрастови групи 11, 13 и 16 години в градовете са систематично по-ниски от претеглените и калиброваните данни. Най-високи тегла са получени при обикновеното претегляне и в трите области за градовете за 11- и 13-годишните в сравнение с непретеглените и калиброваните данни. За учениците на 16 години теглата от простото претегляне заемат междинни стойности между получените относителни дялове от анкетното изследване и калибрираните данни. За посочената последна възрастова група от градовете изчислените калибровъчни тегла имат най-високи стойности.

Разгледаните съотношения за стратите в селата са реципрочни. С най-големи относителни дялове за всички страти са получените от непретеглените данни. Междинно положение по стойност заемат теглата, получени в резултат на калибрацията за възрастовите страти 11 и 13 години и в трите области. За 16-годишните в селата калибровъчните тегла са с най-ниска стойност. За тази възрастова група теглата от простото претегляне имат междинни стойности спрямо теглата, получени от анкетата и калибрацията, а тези от простото претегляне са най-ниски за възрастовите групи 11 и 13 години.

Направените сравнения показват, че при изпълнение на извадката в градовете не е постигнат проектираният обхват и респективно в селата се получава надценяване. Тези несъответствия чрез корекция на теглата за всяка от 18-те страти най-добре са решени чрез получените калибрационни тегла. Предимството на калибрацията се заключава и в това, че присвоените индивидуални тегла на всяка единица по калибровъчните признаци едновременно дава възможност да се произвеждат аналитични таблици, без да е необходимо за всяка таблица да се извършва претегляне. Това е възможно, тъй като всяка от единиците в извадката е претеглена едновременно по всички използвани за калибрацията признаци.

В табл. 3 и 4 са представени данни за разпределението на изследваните лица по признаците пол и възраст.

За признака „пол” не е налична предварителна официална статистическа информация и той не е използван при проектиране на извадката. След като са калибрирани данните по този признак, по-нататък произвежданите таблици ще включват и отразяват и половата структура на изследваната генерална съвкупност.

3. Двумерно разпределение на изследваните лица по пол и възраст

(Проценти)

Пол	Възраст - години			Общо
	11	13	16	
		Непретеглени данни		
Момичета	33.6	37.7	28.8	100.0
Момчета	31.3	29.3	39.5	100.0
Общо	**32.5**	**33.6**	**34.0**	**100.0**
	Тегла, коригирани с неотговорилите (nonrespons)			
Момичета	33.4	37.3	29.3	100.0
Момчета	30.5	28.5	41.0	100.0
Общо	**32.0**	**33.0**	**35.0**	**100.0**
		Калибрирани тегла		
Момичета	36.5	33.7	29.9	100.0
Момчета	34.2	26.8	38.9	100.0
Общо	**35.4**	**30.4**	**34.3**	**100.0**

При сравняването на данните от табл. 3 може да се установи ефектът от допълнителните процедури по претеглянето на данните. В първата част на таблицата е представено полученото от анкетата разпределение на изследваните лица по двата признака - пол и възраст. Във втората част са претеглените резултати от изследването, като са коригирани теглата въз основа на информацията за възрастовите групи и са елиминирани неотговорилите. Фактически тук е отчетено разпределението на изследваните единици в генералната съвкупност само по признака „възраст". Получените тегла в резултат на калибрацията (в третата част) имат поголямо различие спрямо тези от непретеглените и претеглените данни, но тук по-важното е, че едновременно са отчетени разпределенията на изследваните лица в трите области по пол и възрастови групи.

Подобен анализ може да се направи и при получаване на разпределението на изследваните лица по възрастови групи и пол. По този начин ще се отрази половата структура във всяка възрастова група.

4. Двумерно разпределение на изследваните лица по възраст и пол

(Проценти)

Пол	Възраст - години			Общо
	11	13	16	
		Непретеглени данни		
Момичета	53.2	57.7	43.6	51.4
Момчета	46.8	42.3	56.4	48.6
Общо	**100.0**	**100.0**	**100.0**	**100.0**
	Тегла, коригирани с неотговорилите (nonrespons)			
Момичета	53.3	57.7	42.6	51.0
Момчета	46.7	42.3	57.4	49.0
Общо	**100.0**	**100.0**	**100.0**	**100.0**
		Калибрирани тегла		
Момичета	53.2	57.3	45.1	51.7
Момчета	46.8	42.7	54.9	48.3
Общо	**100.0**	**100.0**	**100.0**	**100.0**

При прилагането на процедурите за калибрация на данни от извадки е предвидено, ако получените резултати след калибрацията не удовлетворяват определни математически критерии, по преценка на изследователя е възможно процедурата да се приложи многократно, докато се избере подходящият модел, т.е. възможно е да се отстранят отдалечените стойности на g теглата спрямо границите на определен интервал, който може да се определя напр. като се използват медианната оценка и нейното квартилно отклонение или други критерии. Следователно може да се каже, че калибрацията на данните от извадковите изследвания е един оптимизационен процес за разлика от останалите методи за просто претегляне на данни от извадки.

Математическата сложност на моделите за калибриране на данни изисква използването на специализиран софтуер. Това в определена степен ограничава нейното приложение. Освен това, все още е много малък опитът в нашата страна в областта на калибрирането на данни

при провеждане на извадковите изследвания и специално в областта на социалните изследвания. Все пак първи опити в това отношение има при обработката на данните в някои изследвания на Националния статистически институт, които са включени и в Европейската статистическа система (Евростат).

Прилагането на различните техники за претегляне и преди всичко калибрацията на данните в максимална степен решават проблемите в ситуациите на липсващи данни в Евростат, предизвикани от непълен обхват на планираните извадки, но трябва да се има предвид, че те невинаги могат да компенсират направените пропуски в отделните фази на извадковите изследвания. Във връзка с това не бива да се пренебрегва следователно и прилагането на различните методични и организационно-технически мерки и дейности за преодоляване на пропуските в проектираните извадкови съвкупности.

ЦИТИРАНА ЛИТЕРАТУРА:

Калчев, Й. (2005). Основни проблеми при проектиране на представителни извадкови изследвания, Социологически траектории, УИ „Св. Кл. Охридски“, С., с. 70.

Кокрен, У. (1976). Методы выборочного исследовании, М. с. 381.

ООН, Статистический отдел (2007). Руководство по подготовке статистических данных об использовании времени для оценки оплачиваемого и неоплачиваемого труда, (ST/ESA/STAT/SER.F/93), Нью-Йорк, с. 159 - 160.

Степанов, С. В. (2009). Калибровка выборки, М., с. 1.

Цонев, В. (1970). Основи на репрезентативното изучаване, (литопечат, второ издание), С., с. 74.

Deville, J. C., C. E. Sarndal (1992). Calibration Estimators in Survey Sampling, Journal of the American Statistical Association; p. 376 - 382.

Ekholm, A., S. Laaksonen (1991). Weighting via response modeling in the Finnish Household Budget Survey, Journal of Official Statistics, Sweden.

Kish, L. (1965). Survey sampling, N.Y. Wiley.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. Survey Methodology, p. 133 - 142.

Sarndal, C. E. (2007). The calibration approach in survey theory and practice, Statistics Canada, Vol. 33, No. 2, p. 99 - 119.

Task Force on the implementation of NACE, Rev. 2 (2006). Handbook on methodological aspects related to sampling designs and weights estimations, Eurostat, Annex A, p. 16.

The Quality Guidelines (fourth edition) of Statistics Canada (2003).

КАЛИБРОВКА ДАННЫХ С СОЦИАЛЬНЫХ ИССЛЕДОВАНИЙ

Йордан Калчев, Вера Велева***

РЕЗЮМЕ Значительной проблемой при проведении представительных выборочных обследований является неохват всех единиц запланированного объема выборок по различным причинам. Эти пробелы приводят к нарушениям репрезентативности и точности эмпирических данных. Организационные и методические приемы и инструменты, используемые конструкторами выборочных обследований, не всегда в состоянии компенсировать потерю информации в результате необследованных единиц.

На основе этого тезиса, в статье прослеживается воздействие неохвата на точность получаемых статистических оценок и поддерживается обосновка о необходимости взвешивания получаемых результатов в целях уменьшения размера нестохастической ошибки. В качестве решения проблемы подчеркиваются преимущества разработанных современных статистико-математических моделей для калибровки выборочных данных. На основании специализированной литературы коротко представлена сущность калибровки как новый систематический взгляд в теории и практике выборочных обследований и возможности для получения объективных оценок в случае невыполнения выборок, и даже при других ошибках, связанных с выборками.

Применение калибровки иллюстрируется конкретным примером, основывающимся на данных с конкретного исследования, с использованием продукта g-CALIB, базированного в SPSS.

В заключении подчеркивается, что калибровка, в отличии от других методов взвешивания, являетсяется процессом оптимизации, и эту особенность необходимо учитывать при использовании этой процедуры.

* Д-р, доцент на кафедре социологии Югозападного университета имени Неофита Рильского; e-mail: ikaltchev@abv.bg.

** Главный ассистент на кафедре социологии Югозападного университета имени Неофита Рильского; e-mail: veleva_v@abv.bg.

CALIBRATION OF DATA FROM SOCIAL SURVEYS

Jordan Kaltchev, Vera Veleva***

SUMMARY A significant problem in the implementation of representative sample surveys is not covering due to various reasons of all units in the planned sample sizes. These gaps lead to violations of the representativeness and accuracy of the empirical data. The undertaken by the designers of sample surveys organizational and methodological techniques and tools cannot always compensate for the loss of content as a result of the unexplored units.

Based on this argument, the paper monitors the influence of lack of range in the accuracy of statistical estimates received and the rationale for the need to weigh the results obtained is supported in order to reduce the amount of non-stochastic error. As a solution of the problem the advantages of the developed modern statistical and mathematical models for calibration of data samples are highlighted. Based on the specialized literature, the nature of the calibration system is briefly presented as a new system review in the theory and practice of sample surveys and opportunities to achieve objective assessments whenever there is default and even some other sampling errors.

Application of calibration is illustrated by a specific example based on data from a study using g-CALIB, based in SPSS.

In conclusion it is stated that calibration is an optimization process, unlike other methods of weighting and this feature should be taken into account when using this procedure.

* PhD, Associate Professor, South-West University 'Neofit Rilski'; e-mail: ikaltchev@abv.bg.
** Chief Assistant, South-West University 'Neofit Rilski'; e-mail: veleva_v@abv.bg.