УДК 657

Гладун О.М., к.е.н., старший науковий співробітник,
Інститут демографії та соціальних досліджень
НАН України
Ващаева Н.А., молодший науковий співробітник
Інституту демографії та соціальних досліджень
НАН України

ДИЗАЙН ВИБІРКИ ТА ПРОБЛЕМИ ЙОГО ОПТИМІЗАЦІЇ

В статті розглядаються підходи до побудови вибірки оптимального дизайну, а також питання оптимізації діючого дизайну вибірки. Розглянуті також перелік обмежень, які впливають на результати оптимізації, та підходи до оцінки ефективності оптимізації

Вступ. З розвитком суспільства держава змушена все більше і більше уваги приділяти вирішенню різних соціальних проблем. Все складнішими стають проблеми щодо житлових умов, отримання відповідного медичного обслуговування, умов і оплати праці, набуття освіти, міграційної політики тощо. Розробка відповідних державних та регіональних програм і ефективна їх реалізація неможлива без отримання відповідної інформації від населення. Практично єдиним способом отримання такої інформації є проведення опитувань населення на базі вибіркових обстежень.

Проведення вибіркових обстежень населення в Україні все більше поширюється. Разом з розповсюдженням проведення різноманітних соціологічних та маркетингових досліджень, все більше і більше вибіркові обстеження використовуються органами державного управління. Передові позиції у цьому напрямку займає Держкомстат України, який є головним виробником офіційної статистичної інформації в країні.

Так, органами державної статистики на постійній основі проводяться три вибіркові обстеження населення: умов життя домогосподарств (УЖД), економічної активності населення (ЕАН) та сільськогосподарської діяльності населення у сільській місцевості (СГД). Періодично проводяться різні одноразові обстеження населення з актуальних соціальних проблем. За результатами цих обстежень здійснюється дослідження різних аспектів життя та діяльності населення України. Зокрема, це доходи та витрати населення, рівень освіти, зайнятість, виробництво сільськогосподарської продукції тощо.

Постановка задачі. Проте при використанні даних виникає ряд проблем, обумовлених протиріччям самого вибіркового методу. Це протиріччя полягає, з одного боку, у обмеженості ресурсів на проведення обстеження та, з другого, - у необхідності забезпечення відповідної статистичної надійності оцінок отриманих показників.

Ця проблема гостро стоїть для таких адміністративних територій, як область, район, місто тощо, особливо якщо інформацію необхідно отримати по певних групах населення (наприклад, для сімей з дітьми, пенсіонерами тощо). У цих випадках у групу може потрапити невелика кількість одиниць спостереження і при значній варіативності значень ознаки, надійність отриманих оцінок показників, скоріше за все, буде низька.

Аналіз ситуації щодо надійності окремих даних обстеження УЖД для регіонального рівня вперше проводився ще у 2000 році. У якості критерію надійності використовувався коефіцієнт варіації: якщо його значення було менше 10% показник вважався надійним. Результати аналізу засвідчили, що з 3564 показників щодо витрат населення на продовольчі товари по всіх регіонах тільки 10% були надійними, з 15390 показників щодо витрат населення на непродовольчі товари надійними були тільки 2%, а з 1917 показників щодо витрат на послуги – 4% [1]. Моніторинг надійності даних цього обстеження засвідчує, що приблизно така ж ситуація зберігалась і в наступні роки проведення обстеження.

Малий відсоток надійних показників пояснюється тим, що вибірка для обстеження УЖД розроблялась з метою отримання надійних оцінок основних показників на державному рівні. Низький рівень надійності оцінок окремих показників на регіональному рівні існує і для обстежень ЕАН та СГД. Однак актуальність програм спостереження, потреба в інформації в регіональному та районному розрізах з метою моніторингу стану соціальної та економічної ситуації, призвела до необхідності мати статистично надійну інформацію принаймні на регіональному рівні.

Можна виокремити два напрями вирішення зазначеної проблеми: перший – оптимізація (корегування) дизайну вибірки; другий – застосування статистико – математичних моделей. Огляд моделей найбільш повно зроблений у [2]. При чому використання одного напряму зовсім не виключає використання іншого, можливим (навіть бажаним) є їх комбіноване використання.

Жоден з напрямів не гарантує стовідсоткового результату. Корегування дизайну вибірки ґрунтується на реальних даних, але в значному ступені воно буде базуватись на певних припущеннях та модельних розрахунках. Отримані результати після практичної реалізації вибірки у підсумку також можуть не задовольняти вимозі щодо рівня надійності даних. Розробка моделей є дуже трудомісткою роботою, при чому вона не завжди може бути вдалою. Крім того, сама модель залежить від різних чинників, таких як тип показника, характер даних, наявність додаткової інформації тощо. Проте відсутність гарантії не означає, що ці напрями не слід використовувати.

Ця стаття присвячена першому напряму підвищення надійності даних вибіркових обстежень - оптимізації дизайну вибірки.

Результати дослідження. Спочатку розглянемо поняття “дизайн” та „оптимізація” з точки зору побудови вибірки.

У більшості випадків під дизайном вибірки розуміється логічна модель структури вибіркової сукупності, перелік етапів та принципів її формування. В окремих випадках під дизайном вибірки розуміється конкретна реалізація вибірки. Тобто не тільки модель побудови вибірки, а й вибіркова сукупність, яка отримана відповідно розробленої моделі, та її окремі характеристики.

Дизайн вибірки (та його реалізація на практиці) впливає на надійність даних і кількісно характеризується значенням відносного показника, який отримав назву дизайн-ефект.

Дизайн-ефект показує відмінності варіації оцінок певного показника при реальному дизайну вибірки, від дисперсії оцінок цього ж показника за умови побудови вибірки за принципами простого випадкового відбору, і визначається за формулою [3, 4]:

$$deff(\hat{Y}) = \frac{\hat{V}(\hat{Y})}{\hat{V}_{srs}(\hat{Y})}, \qquad (1)$$

де deff ($\hat{Y}$) - дизайн-ефект для показника $\hat{Y}$;

$\hat{V}(\hat{Y})$ - оцінка варіації показника $\hat{Y}$ для реальної вибірки;

$\hat{V}_{srs}(\hat{Y})$ - оцінка варіації показника $\hat{Y}$ для вибірки такого ж обсягу за умови її побудови за процедурою простого випадкового відбору (simple random sampling).

Вдалий дизайн вибірки дозволяє сформувати оптимальну вибірку в плані надійності даних. Для конкретної реалізації вибірки дизайн-ефект може бути розрахований за наступною формулою [5]:

$$deff = n \cdot \frac{1}{\sigma^2_{srs}} \cdot \left(\frac{CV \cdot \bar{y}}{100\%} \right)^2, \qquad (2)$$

де n – обсяг вибірки;

σ^2_{srs} – дисперсія показника для простого випадкового відбору;

CV – коефіцієнт варіації;

$\bar{y}$ – середнє значення величини показника, що оцінюється.

Таким чином, дизайн-ефект пов’язаний з чотирма змінними. Для конкретної реалізації вибірки можна сказати, що зв’язок між дизайн-ефектом та іншими показниками є функціональним. Парадоксальним є те, що, виходячи з формули (2), зі збільшенням обсягу вибірки значення дизайн-ефекту теж збільшується. Тому, за умови збільшення обсягу вибірки, зміна значення дизайн-ефекту визначається не стільки абсолютним значення самої зміни, скільки

впливом цієї зміни на показники дисперсії та варіації. При формуванні вибірки значення ознак передбачити неможливо, через що значення середньої, дисперсії та коефіцієнту варіації заздалегідь однозначно визначити неможливо. Тому, якщо реалізувати багато вибірок та окремо розрахувати показники та дизайн-ефект, то зв'язок між ними буде мати стохастичний характер.

Зараз найбільш розповсюдженим способом розрахунку дизайн-ефекту є застосування реплікаційних методів, суть яких полягає у формуванні за певним алгоритмом підвибірок з вже відібраної вибіркової сукупності. За результатами формування підвибірок і розраховується значення дизайн-ефекту [6]. Зауважимо також, що дизайн-ефект розраховується не для всієї вибірки, а для кожного показника окремо.

Тепер перейдемо до розгляду поняття оптимальності.

Слід звернути увагу на відмінність термінів оптимальний дизайн та оптимізація дизайну. Як завжди, найбільш вдалим є загальне визначення: оптимальний – це "який найбільше відповідає певним умовам, вимогам, найкращий із можливих" [7, с. 476]. Оптимізація же є синонімом "покращення", тобто удосконалення вже існуючого.

Із визначення випливає, що повинен існувати певний критерій, відповідно до якого і визначається оптимальність рішення. Проте не слід плутати оптимізацію з мінімізацією чи максимізацією (і, відповідно, критерій оптимізації з критеріями мінімізації чи максимізації). Це не тотожні поняття. Мінімізація (максимізація) якогось показника не означає, що він набуває оптимального значення. Наприклад, мінімізуючи витрати на проведення вибіркового обстеження, ми реально не отримаємо оптимального значення витрат з точки зору мети всього обстеження (наприклад, надійності показників, охоплення території тощо). Тобто, окрім однієї цільової функції мінімізації (максимізації) значення параметра, існують інші показники, значення яких нам необхідно утримати у певних межах чи теж мінімізувати (максимізувати). Таким чином, коли мова йде про оптимізацію, то слід говорити про множину критеріїв або один багатовимірний критерій, який формується з цієї множини одновимірних критеріїв.

Відмінність між мінімізацією (максимізацією) та оптимізацією можна сформулювати наступним чином: метою мінімізації (максимізації) є знаходження екстремуму однієї цільової функції або знаходження значення, яке відповідає одному критерію; метою оптимізації є знаходження прийнятного балансу між різними цільовими функціями або критеріями [8].

З цього випливає, що при мінімізації (максимізації) ми можемо мати тільки одне рішення, при оптимізації – декілька. Тобто при проведенні оптимізації необхідне визначення пріоритетності критеріїв. Тому процес визначення оптимального рішення є ітераційним. У підсумку буде отримане формальне рішення, яке відповідає системі вимог. Воно потребує обов'язкового аналізу фахівцями.

При розробці дизайну вибірки для проведення обстежень населення ми маємо декілька умов і критеріїв. Мінімальний їх перелік наступний:

- мінімізація вартості обстеження;
- максимізація надійності оцінок показників (що може бути задано мінімізацією коефіцієнту варіації вибіркових оцінок показників);
- врахування методів оцінювання показників (пряме, композиційне тощо);
- необхідність розподілу вибірки по території країни (наприклад, охоплення всіх регіонів, сільської місцевості тощо).

У таких випадках, коли необхідно отримати інформацію як стосовно всієї генеральної сукупності, так і якоїсь її частини, як правило, використовуються стратифіковані вибірки. Вона гарантує включення у вибірку підгруп, кращі оцінки в результаті більш ефективної вибіркової стратегії.

Після проведення стратифікації обсяг вибірки розподіляється за стратами. В залежності від принципу розподілу виділяють такі види стратифікації:

- проста пропорційна - стратифікація здійснюється пропорційно розміру страт або їх кількості;
- оптимальна - стратифікація здійснюється в залежності від розміру страти та варіації ознаки;
- строго оптимальна - стратифікація здійснюється в залежності від розміру страти, варіації ознаки та вартості витрат на одну одиницю спостереження у кожній страті.

При простій пропорційній стратифікації розподіл обсягу вибірки здійснюється пропорційно розміру страт:

$$n_i = n\frac{N_i}{N}, \qquad (3)$$

де ni – обсяг вибірки і-тої страти;

n – теоретичний (розрахунковий) обсяг вибірки, розрахований для всієї сукупності;

N – обсяг генеральної сукупності;

Ni – обсяг і-тої страти (при цьому $\sum N_i = N$).

Така стратифікація забезпечує побудову рівнозваженої вибірки, що значно спрощує подальші розрахунки оцінок показників та їх надійності.

При проведенні оптимальної стратифікації крім розміру страти враховується і варіація ознаки, покладеної в основу розрахунку обсягу вибірки. При цьому формула розрахунку обсягу вибірки для страти набуває вигляду:

$$n_i = n\frac{N_i\sigma_i}{\sum N_i\sigma_i}, \qquad (4)$$

де σ_i - середньоквадратичне відхилення у і-тій старті.

Такий розподіл обсягу вибірки за стратами ще називають нейманівським розміщенням за прізвищем Є.Неймана, який опублікував цей метод у 1934 році. Але заради справедливості слід зауважити, що першим запропонував цей метод російський статистик О.О.Чупров у 1923 році [9, с.8, 113].

За строго оптимальної стратифікації крім розміру страти та варіації ознаки розподіл загального обсягу вибірки враховує також і вартість витрат на одну одиницю спостереження у кожній страті (за умови незмінності суми загальних витрат):

$$n_i = n\frac{N_i\sigma_i/\sqrt{c_i}}{\sum N_i\sigma_i/\sqrt{c_i}}, \qquad (5)$$

де сі – витрати на одну одиницю вибірки у і-тій страті.

Аналіз цієї формули дає можливість зробити наступні висновки. Обсяг вибірки буде більше для тієї страти, у якої: а) більший розмір; б) більша варіація ознаки; в) менша вартість обстеження одиниці вибірки.

Формула (5) має узагальнюючий характер: при сі = const отримуємо формулу (4) для оптимальної стратифікації, а якщо при цьому і σ_i = const, то отримуємо формулу (3) для простої пропорційної стратифікації.

У Держкомстаті при формуванні вибірок для обстежень УЖД та ЕАН використовувався перший вид стратифікації. Проте, як зазначалось раніше, надійність оцінок основних показників є незадовільною. Тому виникла потреба оптимізації дизайну вибірки.

Проведення оптимізації можливе за наступними варіантами: 1) збільшення обсягу вибірки; 2) зміна територіальної вибірки; 3) перерозподіл вибірки.

При застосуванні першого варіанту можливе загальне (наприклад, пропорційне для всіх регіонів) збільшення обсягу вибірки або тільки для регіонів з недостатньою надійністю даних. При другому варіанті можливе збільшення кількості територій, які обстежуються у регіоні, одночасно зі зменшенням кількості респондентів у кожній території. Третій варіант передбачає збільшення обсягу вибірки у регіонах з недостатньою надійністю даних за рахунок зменшення обсягу вибірки у регіонах, в яких надійність даних висока (іншими словами проведення оптимальної стратифікації). Застосування строгої оптимізації потребує визначення вартості обстеження одного домогосподарства у кожній страті. Можлива і комбінація запропонованих варіантів, наприклад, перерозподіл вибірки одночасною зі зміною територій.

Проте, слід враховувати наступне. Збільшення обсягів вибірки потребуватиме відповідного збільшення обсягів фінансування і термінів обробки отриманої первинної інформації. Додаткового фінансування вимагатимуть практично всі роботи з проведення обстеження: розширення мережі інтерв'юерів, транспортні витрати, друкування інструментарію тощо. Збільшення термінів виконання робіт буде залежати від якості заповнення інструментарію і кількості помилок.

Враховуючи постійне обмеження державної статистики у фінансових ресурсах, при оптимізації мова йде про перерозподіл вибірки.

При перерозподілі вибірки для реальної ситуації, яка є при організації зазначених вибіркових обстежень в Держкомстаті може виникнути ще одна проблема. У зв'язку з тим, що на одних і тих же територіях один інтерв'юер проводить два обстеження у міських населених пунктах та три у сільській місцевості, то може виникнути ситуація, коли по всіх обстеженнях обсяг вибірки буде одночасно збільшений або зменшений. Це викличе проблему щодо навантаження одного інтерв'юера, а також необхідність узгодження навантаження різних інтерв'юерів, що викличе проблеми щодо врегулювання оплати їх праці.

Ще одна проблема полягає у тому, що в залежності від показників, які визначені в якості основних, напрямок перерозподілу вибірки може носити різноспрямований напрямок. Це стосується як різних показників у межах одного обстеження, так і однакових показників різних обстежень.

Таким чином загальна модель оптимізації дизайну вибірки є доволі складною задачею. Тому сформулюємо низку вимог та обмежень, необхідних для розробки моделі.

1) Незмінність загального обсягу вибірки, тобто: $\sum_{i=1}^{H} n_i = n$.

2) Мінімізація варіації вибіркових оцінок показників ($V_i \to \min$).

3) Узгодження зміни обсягів вибірки у стратах за декількома показниками.

4) Узгодження зміни обсягів вибірки у стратах для декількох вибіркових обстежень.

5) Врахування процедури отримання оцінок показників.

6) Визначення як бажаної надійності оцінки кожного показника, так і мінімально прийнятної.

7) Обмеження загальної вартості робіт.

8) Врахування принципів роботи мережі інтерв'юерів.

Окремі вимоги мають протиріччя і потребують відповідного узгодження. Наприклад, обмеження вартості робіт та мінімізація варіації вибіркових оцінок показників.

Після теоретично визначеного оптимального дизайну постає проблема визначення наскільки новий дизайн кращий за існуючий. На практиці це можливо тільки після проведення обстеження на вибірці нового дизайну і порівняння дизайн-ефектів за двома дизайнами вибірки:

$$eff = \frac{deff_N(\hat{Y})}{deff_O(\hat{Y})}, \qquad (6)$$

де eff – ефективність нового дизайну вибірки;

$deff_N(\hat{Y})$ - дизайн-ефект показника $\hat{Y}$ за новим дизайном вибірки;

$deff_O(\hat{Y})$ - дизайн-ефект показника $\hat{Y}$ за старим дизайном вибірки.

Проте бажаною є оцінка ефективності нового дизайну до проведення обстеження. Теоретично це може бути зроблено шляхом розрахунку дизайн-ефекту показника $\hat{Y}$ на базі формування модельної вибіркової сукупності для нового дизайну вибірки.

З формули (6) випливає, що у випадку eff < 1 новий дизайн є кращим за старий і навпаки у випадку eff > 1 – гіршим.

Ефективність може бути оцінена і на підставі розробки аналітичних формул. Для прикладу розглянемо формулу для оцінки ефективності зміни дизайну вибірки за таких умов:

- обсяги вибірок за старим та новим дизайном однакові;
- дисперсії ознак у межах страт однакові;
- стара вибірка була побудована за умови простої пропорційної стратифікації;
- нова вибірка розроблена для умови строгої оптимальної стратифікації.

За наведених умов і при неврахуванні поправки на скінченність сукупності відношення дисперсії при непропорційному розподілі вибірки до дисперсії при пропорційному розподілі може бути представлено у вигляді:

$$eff = \frac{V_{nprop}(\bar{y})}{V_{prop}(\bar{y})} = \frac{\sigma_w^2 \cdot \sum_{h=1}^{H} \frac{W_h^2}{n_{h(nprop)}}}{\frac{\sigma_w^2}{n}} = \frac{\sigma_w^2 \cdot \frac{1}{n^2} \cdot \sum_{h=1}^{H} \frac{n_{h(prop)}^2}{n_{h(nprop)}}}{\frac{\sigma_w^2}{n}} = \frac{\sum_{h=1}^{H} \frac{n_{h(prop)}^2}{n_{h(nprop)}}}{n} \quad (7)$$

де $V_{nprop}(\bar{y})$ - варіація вибіркових оцінок при непропорційному розподілі вибірки;

$V_{prop}(\bar{y})$ - варіація вибіркових оцінок при пропорційному розподілі вибірки;

σ_w^2 - дисперсія ознаки y по одиницях вибіркової сукупності;

$W_h = \frac{n_{h(prop)}}{n}$ - вага h-ої страти (h = 1, 2, …, H);

$n_{h(nprop)}$ - обсяг вибірки h-ої страти при непропорційному розподілі;

$n_{h(prop)}$ - обсяг вибірки h-ої страти при пропорційному розподілі;

n - загальний обсяг вибірки;

H – кількість страт.

Розробка аналітичних формул є складною процедурою, яка залежить від конкретних дизайнів вибірок та умов оптимізації.

Висновки. Оптимізація дизайну вибірки є доволі складною проблемою, яка потребує розробки низки умов оптимізації, узгодження умов між собою, проведення оцінки ефективності нового дизайну у порівнянні з діючим. Вирішення цього комплексу проблем повинно базуватись на використанні даних як реальних, так і модельних вибіркових сукупностей.

Розробка моделі оптимального дизайну залежить від конкретних умов. Отримані результати обов'язково повинні бути проаналізовані та оцінені експертами.

ЛІТЕРАТУРА:

1. *Гладун О.М., Саріогло В.Г.* Напрямки підвищення якості даних вибіркового обстеження умов життя домогосподарств та вирішення проблеми „малих територій" // Статистика України. – 2003, №1, с. 4 – 11.

2. *Rao J.N.K.* Small Area Estimation. – New York: A John Wiley & Sons, Inc., 2003. – 313 p.

3. *Verma V.* Workshop on Labour Force Surveys for CIS Countries and the Baltic States run by the International Labour Office, Bureau of Statistics hosted by State Statistics Committee of Ukraine (Kyiv, 14-19 September 1998). Part Three: sample design, data evalution, field operations.

4. *Kordos J.* Development of Household Surveys in Transition Countries – Some Quality Issues // Statistics in Transition, October 2003, Volume 6, Number 2, pp. 307 – 322.

5. *Гладун О.М.* Вплив дизайну вибірки на показники вибіркового обстеження економічної активності населення // Щорічник наук. праць „Проблеми статистики". –К.: НТК статистичних досліджень Держкомстату України. –2004. –Вип. 6. – с. 99 – 109.

6. *Lehtonen R., Pahkinen E.J.* Practical Methods for Design and Analysis of Complex Surveys. - Revised edition. New Delhi: A John Wiley & Sons, Inc., 1996. – 344p.

7. Новий тлумачний словник української мови (у трьох томах) / Укладачі В.В.Яременко, О.М.Сліпушко. – К.: „Аконіт", 2006. Том 2. – 926 с.

8. *Zeleny M.* Optimality and Optimization (in Russian) // http://artkis.ru/optimization.html#metka1.

9. *Кокрен У.* Методы выборочного исследования. – М.: Статистика, 1976. – 440с.