УДК 311.213.2

Гладун О.М., к.е.н., с.н.с.,
Інституту демографії та соціальних досліджень НАН України

ТЕОРЕТИЧНІ ТА ПРАКТИЧНІ АСПЕКТИ ЗАСТОСУВАННЯ МЕТОДУ PPS ПРИ ФОРМУВАННІ ВИБІРКОВИХ СУКУПНОСТЕЙ НАСЕЛЕННЯ

Розглядаються теоретичні основи проведення відбору з ймовірністю пропорційною розміру (PPS). Розглянуті окремі практичні аспекти використання методу: визначення кількості територіальних одиниць для відбору та розрахунок порогу саморепрезентативності

Вступ. При формуванні вибірок існують два способи відбору одиниць сукупності: простий випадковий та систематичний. Застосування систематичного відбору є більш бажаним, що пояснюється наступними чинниками.

1) При систематичному відборі не треба виготовляти фішки, кульки тощо які б відповідали кожній одиниці основи вибірки.

2) Систематичний відбір технічно простіше реалізувати, тому що одиниці вибірки визначаються шляхом розрахунків. До того ж при правильній побудові основи вибірки систематичний відбір практично не відрізняється від простого випадкового. Крім того, при застосуванні певних принципів упорядкування, систематичний відбір може забезпечити кращу репрезентативність вибірки по характеристиках, покладених в основу упорядкування. Упорядкування при систематичному відборі є по суті неявною стратифікацією.

3) Систематичний відбір забезпечує ту ж саме похибку вибірки, що і простий випадковий.

Систематичний відбір, в свою чергу, поділяється на відбір з рівними ймовірностями відбору і з нерівними.

При рівній ймовірності відбору кожна одиниця основи вібірки має рівні шанси бути відібраною. Але не завжди такий підхід узгоджується як з метою обстеження, так і зі звичайною логікою. Наприклад, при проведенні обстеження населення необхідно відібрати міські населені пункти, в яких потім буде опитуватись населення. При рівній ймовірності відбору всі населені пункти – і місто Київ з чисельністю 2693,2 тис. осіб, і місто Керч з чисельністю 151,3 тис. осіб, і місто Свалява з чисельністю 16,9 тис. осіб тощо – будуть мати однаковий шанс потрапити до територіальної вибірки. Навряд чи з таким підходом можно погодитись. Окрім того, якщо кількість міст, які необхідно відібрати невелика, це практично завжди вплине на якість вибірки, і, відповідно, якість отриманих даних. До того ж при цьому значно ускладнюється процедура обробки даних.

Тобто в певних випадках доцільно при відборі враховувати розмір одиниць, які відбираються (у нашому прикладі в якості розміру виступає чисельність населення міста). В таких випадках використовується метод відбору з ймовірностями, пропорційними розміру або за загальноприйнятою англійською абревіатурою *PPS* (probability proportional to size).

Постановка проблеми. Незважаючи на дуже широке його використання, в сучасній українській статистичній літературі теоретичні та особливо практичні аспекти цього методу залишаються практично не висвітленими. Так, в своїй роботі щодо вибіркових обстежень А.М.Єріна присвячує цьому методу кілька сторінок при розгляді процедури формування вибірки навіть не згадуючи назву методу [1, с. 78, 79]. В.Г.Саріогло розглядає цей метод в контексті розрахунку системи ваг при складному дизайні вибірки [2, с.90 – 107]. Складності практичного характеру, які виникають при застосуванні методу *PPS* залишаються не висвітленими та не вирішеними.

Тому **метою дослідження** є розгляд теоретичних засад методу відбору *PPS* та

розробка методичних положень вирішення окремих практичних проблем, які виникають при застосуванні методу.

Викладення основного матеріалу дослідження. Спочатку розглянемо теоретичні аспекти застосування методу PPS.

Нехай населення, яке треба обстежити, проживає на *A* територіальних одиницях, із них треба відібрати *a* одиниць. Чисельність населення *i-тої* територіальної одиниці позначимо як M_i, тоді $\sum_{i=1}^{A} M_i = N$ (*i*– номер територіальної одиниці *(i = 1, ..., A)*; *N* – обсяг генеральної сукупності). Для відібраних територіальних одиниць ймовірність відбору дорівнює:

$$P_j = a\frac{M_j}{\sum_{i=1}^{A} M_i}, \qquad (1)$$

де *j* – номер відібраної територіальної одиниці *(j = 1, ..., a)*.

За умови двоступеневого відбору, на другому ступені з кожної відібраної *j-тої* територіальної одиниці відбираються домогосподарства. Ймовірність відбору *k-ого* домогосподарства в середині *j-тої* територіальної одиниці дорівнює:

$$P_k = \frac{n_j}{M_j}, \qquad (2)$$

де n_j – кількість домогосподарств, яку треба відібрати з *j-тої* територіальної одиниці.

При цьому сума домогосподарств, відібраних у всіх *a* територіальних одиницях дорівнюватиме обсягу вибірки: $\sum_{j=1}^{a} n_j = n$.

Загальна ймовірність відбору домогосподарства становить:

.

(3)

Кількість домогосподарств n_j визначається за формулою (4) виходячи з умови забезпечення однакової ймовірності відбору всім кінцевим одиницям вибірки:

$$\left(a\frac{M_j}{\sum_{i=1}^{A} M_i}\right)\cdot\left(\frac{n_j}{M_j}\right) = \frac{n}{N} = f, \qquad (4)$$

де *f* – частка відбору.

Суть методу *PPS* полягає у тому, що після проведення відбору теоретична ймовірність відбору кожної одиниці вибірки (домогосподарства) повинна бути рівною. Розмір територіальної одиниці як раз і вливає на значення ймовірностей. Відповідно, рівними повинні бути і ваги кожного домогосподарства: $w_h = \frac{1}{P_{jk}}$, де w_h – вага *h-ого* домогосподарства *(h = 1, ..., n)*. При цьому сума ваг всіх домогосподарств повинна дорівнювати обсягу генеральної сукупності: $\sum_{h=1}^{n} w_h = N$.

При більшій кількості ступенів відбору процедура відбору територій залишається без змін. Додаткові ступені (третій, четвертий тощо) можуть виникнути з двох причин: 1) коли після відбору територій відсутня інформаційна база для проведення відбору домогосподарств або створення та використання такої бази занадто трудомістке: 2) представлення менших за розміром територій обумовлено вимогами до вибірки, а проведення відбору відразу менших територіальних одиниць або неможливе з причини відсутності інформаційної бази, або коли метою обстеження передбачається отримання інформації по різних за ієрархією територіальних одиниць. На практиці у більшості випадків вистачає як раз триступеневого відбору.

Так, формування вибірки у міських поселеннях може здійснюватись за наступною схемою: спочатку відбираються міські населені пункти (територіальні одиниці), в середині відібраних міських населених пунктів – локальні територіальні одиниці (виборчі дільниці, поштові відділення, переписні дільниці тощо). По відібраних територіальних одиницях вже можна скласти

перелік адрес домогосподарств і здійснити відбір необхідної їх кількості.

При формуванні вибірки у сільській місцевості доцільно на першому ступені відбирати райони (територіальні одиниці), а в середині районів краще за все відбирати сільські ради (локальні територіальні одиниці). Це пояснюється тим, що в сільській місцевості у межах кожної сільської ради існує погосподарський облік. Дані по господарського обліку дають можливість отримати актуалізовану на початок року інформацію щодо кількості домогосподарств у сільській раді та списки домогосподарств. Це значно спрощує процедуру формування списків та робить точними розрахунки ймовірностей відбору.

Далі розглянемо окремі практичні аспекти використання методу *PPS*.

До початку робіт з формування вибірки неможливо визначити кількість домогосподарств, які необхідно відібрати з кожної територіальної одиниці на останньому ступені відбору. Це пояснюється двома наступними чинниками:

– доволі часто заздалегідь невідома кількість територіальних одиниць, яку необхідно відібрати (параметр a у формулах (1), (3), (4));

– невідомий розмір територіальних одиниць, які сформують територіальну вибірку (параметр M_j у формулах (1) – (4)).

Проблема визначення кількості територіальних одиниць, яку треба відібрати, розглядається далі. Конкретна ж територіальна одиниця та, відповідно, її розмір з'ясовуються безпосередньо під час проведення відбору територій (формування територіальної вибірки).

Якщо ці параметри відомі, то кількість одиниць відбору (домогосподарств) визначається за необхідності забезпечення рівної ймовірності відбору для кожного домогосподарства за формулою (4). Ідеально дотриматись співвідношень, які визначаються цими формулами, на практиці неможливо. Це пояснюється наступними причинами.

По-перше, відбирати на певній територіальній одиниці кількість домогосподарств, меншу за певне число, часто буває недоцільно. Така ситуація може виникнути, коли обсяг вибірки малий, а необхідно забезпечити максимальне територіальне охоплення. У цьому випадку у територіальній вибірці може бути багато територій, які знаходять одна від одної на значній відстані, що значно ускладнює організацію робіт з проведення обстеження.

По-друге, не завжди вдається чітко дотриматись умови $\sum_{i=1}^{A} M_i = N$. Це може бути, наприклад, у випадку, коли генеральною сукупністю є певна група населення, точні дані відносно якої відсутні у розрізі територіальних одиниць. Тоді задовільною є умова, що $\sum_{i=1}^{A} M_i \approx N$.

По-третє, до певних відхилень може призвести і використання різних одиниць виміру територіальних одиниць.

Так, наприклад для триступеневої вибірки, при відборі міських територій, на першому ступеню у якості розміру населеного пункту краще використовувати чисельність населення. Це пояснюється проведенням щорічних розрахунків саме чисельності населення всіх міських населених пунктів, у той час як кількість домогосподарств точно може бути визначена раз на десять років під час проведення переписів населення. На другому ступені відбору розмір локальної територіальної одиниці може бути визначений або через чисельність населення, або через кількість адрес домогосподарств (які виступають в якості оцінки кількості домогосподарств). Чисельність населення в якості розміру може бути використана, коли локальною територіальною одиницею виступають внутрішньоміські райони; а чисельність адрес – при використанні у процедурі відбору поштових відділень, виборчих дільниць тощо. На третьому ступені відбору немає альтернати використання адрес, тому що точну інформацію по локальній

територіальній одиниці вже не складно отримати і тому що адреса все одно потрібна в більшості випадків проведення вибіркових обстежень населення.

У сільській місцевості в багатьох випадках завдяки налагодженому погосподарському обліку на всіх ступенях відбору в якості розміру територіальної одиниці доцільно використовувати кількість домогосподарств.

Відмінності між теоретично визначеними та реально отриманими ймовірностями враховуються на етапі обробки даних за допомогою введення для кожного домогосподарства спеціальних коефіцієнтів.

А тепер розглянемо проблему визначення кількості територіальних одиниць, яку потрібно відібрати.

Тут можливі наступні чотири варіанти: перший – коли кількість територіальних одиниць визначена заздалегідь; другий – коли заздалегідь визначений мінімальний розмір, більше якого всі територіальні одиниці включаються до вибірки; третій – коли кількість територіальних одиниць заздалегідь не визначена, але є визначеним навантаження інтерв'юера; четвертий – коли не визначена ані кількість територіальних одиниць, ані навантаження інтерв'юера.

Спочатку розглянемо коефіцієнт a у формулі (1). Його наявність призводить до того, що територіальні одиниці, які мають розмір, рівний чи більший за певне значення, яке називають „пороговим", мають ймовірність відбору P_j рівну чи більшу одиниці.

При застосуванні методу відбору *PPS* поріг саморепрезентативності і є критерієм визначення територіальних одиниць, які обов'язково мають бути включені до територіальної вибірки: якщо їх розмір більший за поріг саморепрезентативності, то такі територіальні одиниці обов'язково включаються до територіальної вибірки і називаються саморепрезентативними; якщо їх розмір менший за поріг саморепрезентативності, то такі територіальні одиниці відбираються із застосуванням певних процедур відбору і називаються несаморепрезентативними. Порогове значення (або „поріг саморепрезентативності") у цьому випадку визначається за формулою:

$$N_r = \frac{\sum_{i=1}^{A} M_i}{a}, \qquad (5)$$

де N_r – поріг саморепрезентативності.

Формула (5) відповідає першому варіанту визначення порогу саморепрезентативності.

При другому варіанті порогове значення визначається виходячи з мети обстеження. (Наприклад, коли необхідно, щоб обов'язково було обстежене населення всіх міст чисельністю 100 тисяч осіб і більше.) У цьому випадку нам відоме значення N_r , а задача полягає у визначенні коефіцієнт a – кількості територіальних одиниць.

Загальна кількість територіальних одиниць a буде складатись з двох доданків: кількості територіальних одиниць, розмір яких перевищує встановлений поріг саморепрезентативності (a_{mr}), та кількості територіальних одиниць, які необхідно відібрати з числа тих територіальних одиниць, розмір яких менший за поріг саморепрезентативності (a_{sr}):

$$a = a_{mr} + a_{sr} . \qquad (6)$$

Значення a_{mr} визначається шляхом прямого підрахунку кількості таких територіальних одиниць, а значення a_{mr} розраховується за формулою:

$$a_{sr} = \frac{\sum_{i=1}^{A} M_i - \sum_{m=1}^{a_{mr}} M_m}{N_r}, \qquad (7)$$

де M_m – розмір *m-тої* територіальної одиниці більшої за поріг саморепрезентативності $(m = 1, ..., a_{mr})$.

Розглянемо третій варіант – розрахунок порогу репрезентативності виходячи з завантаження одного інтерв'юера.

На початку формування територіальної вибірки для обстежень населення, кількість територіальних одиниць, яку необхідно відібрати, невідома. Тому використовується підхід, коли поріг саморепрезентативності

визначають на основі обсягу вибірки n та кількості домогосподарств (населення), що обстежуються одним інтерв'юером (навантаження інтерв'юера), за умови повного його завантаження при обстеженні даної територіальної одиниці.

Поріг саморепрезентативності визначається за наступною формулою:

$$N_r = \frac{N}{n} \cdot q, \qquad (8)$$

де q – початкове навантаження інтерв'юера (домогосподарств).

Проте, доволі часто розмір території (особливо населеного пункта) визначається у особах. У цьому випадку для розрахунку порогу саморепрезентативності у формулу (8) слід ввести додатковий множник – середній розмір домогосподарства:

$$N_r = \frac{N}{n} \cdot q \cdot q_h, \qquad (9)$$

де q_h – середній розмір домогосподарства (осіб).

Але іноді використання одного з трьох зазначених підходів визначення порогу саморепрезентативності при підготовці вибіркового обстеження є неможливим. Такі випадки можуть бути при проведенні обстеження вперше, коли ще не визначені деякі вимоги до обстеження, зокрема навантаження інтерв'юера. Тому при четвертому варіанті пропонується наступний підхід.

Поріг саморепрезентативності можне визначатись мінімальною чисельністю населення, для якої можуть бути отримані дані заданої надійності відповідно до вимог до побудови вибірки. Для цього можуть бути застосовані формули, які використовуються для розрахунку обсягу вибірки [3]. Наприклад, для частки формула буде мати наступний вигляд:

$$n_r = deff \cdot \hat{\sigma}^2_{srs} \cdot \left(\frac{100\,\%}{CV \cdot \hat{\omega}} \right)^2, \qquad (10)$$

де n_r – обсяг вибірки, для якої можуть бути отримані дані заданої надійності відповідно до вимог до побудови вибірки;

deff – дизайн – ефект (визначається як відношення дисперсії оцінки показника, який вимірюється у вибірковому спостереженні, для реального дизайну вибірки, до дисперсії оцінки цього показника за припущенням побудови вибірки за принципом простого випадкового відбору);

σ^2_{srs} – дисперсія, яка характеризує варіацію значень показника по одиницях вибірки, за умови її побудови за процедурою простого випадкового відбору;

$\hat{\omega}$ – оцінка очікуваного значення частки;

CV – коефіцієнт варіації.

Після визначення n_r, поріг саморепрезентативності розраховується за формулою:

$$N_r = n_r / f. \qquad (11)$$

Проте четвертий підхід дуже чутливий до параметрів формул (10) та (11). При малих значеннях коефіцієнту варіації та частки відбору отримуємо великі порогові значення. Їх можна застосовувати тільки для великих територіальних одиниць, таких як області. При збільшенні значень коефіцієнту варіації та частки відбору цей підхід може бути застосований до міських населених пунктів та районів.

Зазначимо, що цей підхід доцільно використовувати і при рішенні іншої задачі: по яких за розміром територіях будуть отримані дані визначеної надійності. При певних умовах це дасть інформацію для об'єднання територіальних одиниць при проведенні відбору та наступній обробці даних обстеження.

Після визначення порогу саморепрезентативності та кількості несаморепрезентативних одиниць, проводиться відбір цих одиниць. Вони відбираються до вибірки з ймовірністю, пропорційною їх розміру. Несаморепрезентативні територіальні одиниці вибираються з дотриманням наступної умови: одна така одиниця репрезентує групу територіальних одиниць, сума розмірів яких дорівнює порогу саморепрезентативності.

Процедура відбору несамопрезентативних одиниць складається з наступних етапів: упорядкування одиниць відбору; розрахунок накопичених сум; визначення кроку відбору;

визначення першої одиниці для відбору; відбір наступних одиниць. Розгляд самої процедури відбору виходить за межі цієї статті.

При проведенні крупних обстежень в окремих випадках доцільно проводити кластерний відбір, який передбачає обстеження всього населення на відібраній території. Перш за все, це стосується локальних територій, тому що забезпечити суцільне обстеження цілого міста або сільського району дуже важко з причини значних організаційних, фінансових та методологічних проблем. Використання локальних територій в якості кінцевих одиниць відбору може бути обумовлена метою обстеження, наявною інформаційною базою тощо.

Але у цьому випадку використання локальних територій в якості кінцевих одиниць відбору призводить до ситуації, коли до завершення процедури відбору невідома сумарна чисельність населення на всіх відібраних локальних територіях. Це породжує певну невизначеність щодо обсягу вибірки та кількості локальних територій, яку необхідно відібрати. Це залежить від чисельності населення у кожній локальній території. Ймовірність того, що буде точно відібраний розрахований обсяг вибірки, невелика.

Тому в такому випадку пропонується при проведенні відбору переходити від чітко розрахованого обсягу вибірки до його інтервальної оцінки. Суть полягає у тому, що визначається певний інтервал чисельності населення, у який повинна потрапити сумарна чисельність відібраного населення. Відбір локальних територій проводиться послідовно за певним алгоритмом і припиняється, коли сумарна чисельність населення по відібраних локальних територіях потрапить у визначений інтервал.

Висновки та перспективи подальших досліджень. Таким чином, відбір з ймовірністю пропорційною розміру є процедурою, яка забезпечує кожний одиниці генеральної сукупності рівну ймовірність бути відібраною. Запропоновані методичні положення вирішення практичних проблем реалізації метода *PPS*, забезпечують його коректну реалізацію, що сприяє отриманню якісної вибірки.

Список використаної літератури:

1. *Єріна А.М.* Організація вибіркових обстежень. – К.: КНЕУ, 2004. – 127 с.

2. *Саріогло В.Г.* Проблеми статистичного зважування вибіркових даних. – К.: ІВЦ Держкомстату України, 2005. – 264 с.

3. Методологічні основи формування вибіркових сукупностей для проведення органами державної статистики України базових державних вибіркових обстежень населення (домогосподарств) / Затверджені наказом Держкомстату України від 2 серпня 2005 р. – № 223. – К., 2005 р. – 40 с.

ГЛАДУН Олександр Миколайович – кандидат економічних наук, старший науковий співробітник, провідний науковий співробітник Інституту демографії та соціальних досліджень НАН України

Наукові інтереси:
– вибіркові обстеження населення;
– соціально-демографічна статистика;
– обробка та аналіз даних.

Дата надходження статті 27 березня 2007 року.