Методологія обстежень:
Відбір – калібрація – оцінка

Маркус Гінтас Шова[1] та Каміль Вандерхофт[2]

Частина I: Теорія

Маркус Гінтас Шова

Київ – Kyiv 15-19.XI.2004

[1]ONS – Офіс національної статистики, Великобританія
e-mail: markus.sova@ons.gov.uk

[2]NSI – Національний Статистичний Інститут, Бельгія
VUB – Вільний Брюссельський університет, Бельгія
e-mail: Camille.Vanderhoeft@statbel.mineco.fgov.be

[1] ONS – Офіс національної статистики, Великобританія
[2] VUB – Вільний Брюссельський університет, Бельгія
email: Camille.Vanderhoeft@statbel.mineco.fgov.be

1. Загальні поняття відбору та оцінки на основі дизайну

1.1 Вступ та принципи

Підхід до обробки результатів обстежень, оснований на дизайні, базується на принципі, за яким кожний елемент генеральної сукупності має фіксовані значення обстежуваних змінних. Є кілька причин, що можуть спонукати нас провести вибіркове, а не суцільне, обстеження певної сукупності. Головним аргументом виступають витрати. Суцільне обстеження є дорогим та ресурсомістким. Крім того, добре побудована вибірка може дати точніший результат за суцільне обстеження, оскільки проблеми якості даних можна виправляти лише для вибіркових, а не суцільних обстежень.

З суті відбору випливає, що різні вибірки з однієї й тієї ж генеральної сукупності дають різні оцінки показників сукупності. Оскільки значення сукупності є фіксованими, необхідно скористатися певним випадковим процесом, аби оцінки мали статистичні властивості. Тому ми використовуємо метод відбору на основі ймовірності.

Цей курс включає ряд методів відбору та оцінки для підходів на основі дизайну та на основі моделі. Більш детальне обговорення цих методів обробки обстежень міститься у Särndal *et al.* (1992) та Cochran (1977). Ці роботи вважаються класичними у цьому напрямку.

1.2 Простий випадковий відбір

Простий випадковий відбір (ПВВ) – найпростіша форма ймовірнісного відбору. З сукупності, що складається з N елементів, ми вибираємо один елемент, причому ймовірність вибору кожного елементу сукупності дорівнює $1/N$. Далі ми вибираємо другий елемент з решти сукупності, причому ймовірність вибору кожного елементу дорівнює $1/(N\text{-}1)$. Далі ми продовжуємо відбір таким самим чином, поки не отримаємо вибірку з n елементів. Тоді кожний елемент вибірки має ймовірність включення n/N, а кожна вибірка з n елементів має однакову ймовірність бути вибраною.

Якщо ми хочемо оцінити середнє значення змінної y для сукупності, ми використовуємо середнє для вибірки як оціночну функцію:

$$\hat{\bar{y}}_U = \bar{y}_s = \frac{1}{n}\sum_{i\in s} y_i \tag{1.1}$$

де U – генеральна сукупність, s - вибірка, а y_i – значення обстежуваної змінної для елементу i.

Підсумкове значення для сукупності t_y є добутком середнього значення та розміру сукупності N. Можна оцінити підсумок, підставивши середнє у наступне співвідношення:

$$\hat{t}_y = N\hat{\bar{y}}_U = \frac{N}{n}\sum_{i\in s} y_i \tag{1.2}$$

Рівняння 1.2 – це зважена сума вибіркових значень, причому вага значення для кожного елементу дорівнює N/n. Ця вага – *вага дизайну,* або *вибіркова вага*. Вона залежить від методу відбору елементів вибірки.

Якщо ми бажаємо оцінити долю сукупності, яка має певну характеристику, можна вважати кожне y_i як показник для елементу i. Іншими словами, y_i приймає значення 1, якщо елемент i має характеристику, або 0, якщо i не має цієї характеристики. Оцінка шуканої долі перетворюється на оцінку середнього значення обстежуваної змінної для сукупності, яку можна отримати за рівнянням 1.1.

Дисперсія сукупності визначається так:

$$\begin{aligned} S_{yU}^2 &= \frac{1}{N-1}\sum_{i\in U}\left(y_i - \bar{y}_U\right)^2 \\ &= \frac{\sum_{i\in U} y_i^2}{N-1} - \frac{\left(\sum_{i\in U} y_i\right)^2}{N\left(N-1\right)} \end{aligned} \tag{1.3}$$

Її можна оцінити через вибіркову дисперсію:

$$\begin{aligned} S_{ys}^2 &= \frac{1}{n-1}\sum_{i\in s}\left(y_i - \bar{y}_s\right)^2 \\ &= \frac{\sum_{i\in s} y_i^2}{n-1} - \frac{\left(\sum_{i\in s} y_i\right)^2}{n\left(n-1\right)} \end{aligned} \tag{1.4}$$

З цього можна вивести дисперсію оцінюваного підсумку для сукупності:

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{t}_y\right) = N^2 \frac{\left(1-f\right)}{n} S_{ys}^2 \tag{1.5}$$

а також дисперсію оцінюваного середнього або долі у сукупності:

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\bar{y}}_U\right) = \frac{\left(1-f\right)}{n} S_{ys}^2 \tag{1.6}$$

У формулах 1.5 та 1.6 $f=n/N$ – це *відносний розмір вибірки,* тобто частка сукупності, що охоплюється вибіркою.

1.3 Стратифікований простий випадковий відбір

Інколи можливо розділити генеральну сукупність на підсукупності за якимось критерієм (який дозволяє віднести кожний елемент до однієї й лише однієї підсукупності). Якщо елементи в рамках підсукупності є в чомусь подібними, а елементи з різних підсукупностей є менш подібними, ми можемо скористатися цією інформацією для покращення відбору та оцінки. Якщо відбір у кожній підсукупності ведеться незалежно від відбору з з інших підсукупностей, ми називаємо такі підсукупності *стратами.* Нейман (1934) описує структуру стратифікованого простого випадкового відбору (СПВВ – тобто простого випадкового відбору, що ведеться окремо у кожній страті), відкидаючи ідею "загально репрезентативної вибірки" та аргументуючи на користь ідеї "репрезентативного методу відбору".

При стратифікації підсумок за сукупністю можна розглядати як суму підсумків за стратами. Тому при СПВВ з H стратами ми отримуємо наступний вираз оціночної функції (1.2) для підсумку за сукупністю, де h – індекс окремих страт:

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_h} y_i \tag{1.7}$$

Слід відзначити, що вага дизайну N_h/n_h може бути різною для кожної страти, проте вона є ідентичною для кожного відібраного елементу в рамках однієї страти. Незалежність відбору у різних стратах дає нам наступну оцінку дисперсії на основі рівняння (1.5):

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{t}_y\right) = \sum_{i=1}^{H} \hat{\mathrm{V}}\mathrm{ar}\left(\hat{t}_{yh}\right) = \sum_{h=1}^{H} N_h^2 \frac{\left(1-f_h\right)}{n_h} S_{ys_h}^2 \tag{1.8}$$

Використання стратифікації вимагає відповіді на питання про розподіл вибірки між стратами. Іншими словами, маючи загальний розмір вибірки n, якими мають бути розміри підвибірок n_h для кожної страти? Один простий метод – розподілити вибірку між стратами пропорційно до розмірів цих страт:

$$n_h \propto N_h$$
$$i.e.\quad n_h \approx N_h \frac{n}{N} \tag{1.9}$$

Ця тотожність є приблизною, оскільки розміри підвибірок зі страт мають бути цілими числами, в той час як у правій частині рівняння (1.9) може й не бути цілою. Цей метод називають *пропорційним розподілом*. Нейман (1934) показав, що якщо дисперсії сукупностей у стратах відомі, оптимальним є наступний розподіл:

$$n_h \propto N_h \sqrt{S^2_{yU_h}}$$
$$i.e.\quad n_h \approx n \frac{N_h \sqrt{S^2_{yU_h}}}{\sum_{k=1}^{H} N_k \sqrt{S^2_{yU_k}}} \tag{1.10}$$

Під "оптимальним" ми розіміємо розподіл, який забезпечує найменшу дисперсію оціночної функції підсумку для всієї сукупності. Такий розподіл називають *розподілом Неймана.* Хоча на практиці малоймовірно, що дисперсії сукупностей у стратах відомі, проте за даними минулих обстежень їх можна приблизно оцінити.

1.4 Відбір з неоднаковою ймовірністю

У розділі 1.2 ми бачили, що при ПВВ кожний елемент сукупності має однакову ймовірність відбору, тобто *ймовірність включення* до вибірки. У розділі 1.3 ми побачили, що при СПВВ елементи сукупності можуть мати неоднакову ймовірність відбору, якщо вони належать до різних страт. Горвіц та Томпсон (1952) розширили цю ідею, припустивши можливість для кожного елементу мати різну ймовірність включення, яку ми позначимо π_i для елементу i. Це метод *відбору з неоднаковою імовірністю* (ВНІ), у якому для підсумку сукупності використовується *оціночна функція Горвіца-Томпсона*:

$$\hat{t}_{\pi y} = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{1.11}$$

Зазначимо, що кожний відібраний елемент має вагу, зворотну до його ймовірності включення. Більше того, необов'язково знати розмір сукупності *N,* а розмір вибірки *n* може бути випадковою, а не фіксованою величиною.Можна побачити, що формули 1.2 для ПВВ та 1.7 для СПВВ – це спеціальні випадки загальної формули 1.11. Легко показати, що оцінка Горвіца-Томпсона для підсумку сукупності *t* є незміщеною відносно дизайну, якщо кожний елемент сукупності має ймовірність включення більше нуля. Проте це не означає, що це хороша оціночна функція для кожної схеми відбору, наприклад, для випадку слонів Басу (Basu, 1971).

Оцінка Горвіца-Томпсона має наступну дисперсію:

$$\mathrm{Var}\left(\hat{t}_{\pi y}\right) = \sum_{i \in U} \sum_{j \in U} \frac{\left(\pi_{ij} - \pi_i \pi_j\right) y_i y_j}{\pi_i \pi_j} \tag{1.12}$$

яку можна оцінити таким виразом:

$$\hat{\mathrm{Var}}\left(\hat{t}_{\pi y}\right) = \sum_{i \in s} \sum_{j \in s} \frac{\left(\pi_{ij} - \pi_i \pi_j\right) y_i y_j}{\pi_{ij} \pi_i \pi_j} \tag{1.13}$$

Тут π_{ij} – *спільна ймовірність включення* (або *ймовірність включення другого порядку*) елементів *i* та *j*, тобто ймовірність включення до вибірки обох елементів. На жаль, оціночна функція 1.13 може приймати негативні значення, якщо відібрані елементи мають високу ймовірність включення. Тому замість 1.13 часто використовують оціночну функцію Сена-Йейтса-Гранді:

$$\hat{\mathrm{Var}}_{SYG}\left(\hat{t}_{\pi y}\right) = -\sum_{i \in s} \sum_{\substack{j \in s \\ j<i}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{1.14}$$

Оціночна функція Сена-Йейтса-Гранді також може приймати негативні значення, якщо $\pi_i \, \pi_{j<} \, \pi_{ij}$ для якихось $i \neq j$, проте у більшості практичних випадків це не відбувається. Варто зауважити, що ймовірності включення другого порядку (потрібні для оціночних функцій дисперсії 1.13 та 1.14) може бути важко оцінити для певних схем відбору. Більше того, ймовірності включення другого порядку мають бути більше нуля для кожної пари елементів сукупності.

Якщо для певної схеми відбору усі ймовірності включення першого та другого порядку більше нуля, її називають *вимірюваною схемою відбору.*

Хайек запропонував альтернативу оцінці Горвіца-Томпсона, яка зменшує дисперсію ціною більшого зміщення. Оціночна функція Хайека відома як ”циркова оцінка”, оскільки вона є відповіддю на випадок слонів Басу:

$$\hat{t}_{Hy} = N \frac{\sum_{i \in s} \frac{y_i}{\pi_i}}{\sum_{i \in s} \frac{1}{\pi_i}} \qquad (1.15)$$

Як метод відбору з неоднаковою ймовірністю широко використовується *ймовірність, пропорційна до розміру* (PPS). Для цього методу необхідно знати якусь змінну “розміру” для кожного елементу сукупності. Ймовірність включення елементу при цьому є пропорційною до його розміру.

2. Більш складні методи відбору та оцінки

2.1 Кластерний відбір

Можлива ситуація, коли база вибірки (тобто список усіх елементів генеральної сукупності) не існує, проте є список груп елементів генеральної сукупності. Можливо також, що елементи генеральної сукупності належать до компактних груп, розсіяних по великій географічній території. За таких обставин має сенс застосувати схему відбору елементів, що враховує інформацію щодо їх груп. Одна з таких схем – це *кластерний відбір.* Кожний елемент сукупності належить до одного й лише одного кластера, і ми маємо основу вибірки у вигляді кластерів. З цієї основи відбирається ймовірнісна вибірка кластерів, і обстежуються усі елементи відібраних кластерів.

Наведемо приклад кластерного відбору. Нехай ми хочемо провести обстеження щодо розміру класних кімнат у школах України. В нас може бути список усіх шкіл, проте не список приміщень у школах. Ми можемо відібрати для вибірки ряд шкіл і далі виміряти розмір кожної класної кімнати у вибраних школах.

Оскільки обстежуються усі елементи сукупності в рамках відібраних кластерів, ймовірність включення кожного елементу дорівнює ймовірності включення кластера, до якого належить елемент. Для оцінки підсумку для усієї сукупності можна скористатися функцією Горвіца-Томпсона 1.11.

2.2 Двоступеневий відбір

Повертаючися до нашого прикладу обстеження розмірів класних кімнат, уявімо собі, що ми вимірюємо розміри лише випадкової вибірки кімнат у кожній школі (аби менше заважати заняттям). Це приклад *двоступеневого відбору*. При двоступеневому відборі ми маємо справу з *первинними елементами вибірки (ПЕВ)* замість кластерів. Елементи віідбираються до вибірки згідно з певною схемою випадкового відбору. Перший та другий етап відбору відбуваються незалежно один від одного. Тому ймовірність включення елементу генеральної сукупності є добутком ймовірності його включення на другому етапі та ймовірності включення відповідного ПЕВ на першому етапі. Як і у випадку кластерного відбору, можна скористатися функцією Горвіца-Томпсона 1.11 для оцінки підсумку для всієї сукупності. Фактично кластерний відбір – це спеціальний випадок двоступеневого відбору, коли на другому етапі усі ймовірності включення дорівнюють одиниці.

Принципи двоступеневого відбору можна поширити на більше число етапів, проте слід обережно ставитися до розрахунку ймовірностей включення другого порядку для оцінки дисперсії.

2.3 Оцінка для домена

В обстеженнях часто необхідно оцінити певний параметр для однієї чи кількох підмножин (які називають *доменами)* генеральної сукупності. Не завжди можна визначити, які елементи сукупності належать до домену. Наприклад, нехай ми проводимо обстеження росту працівників ДКСУ і хочемо отримати оцінки для доменів, визначених за кольором очей (блакитні, зелені, карі та ін.). У ДКСУ є список працівників, але в ньому навряд чи вказується колір очей. Тому належність до доменів визначається лише після відбору елементів до вибірки та збору відповідних даних. Якщо індекс d позначає домен, оціночна функція Горвіца-Томпсона набуває такого вигляду:

$$\hat{t}_{\pi yd} = \sum_{i \in s_d} \frac{y_i}{\pi_i} \qquad (2.1)$$

Тут необхідно знати розмір домену N_d, який можна оцінити, використоввуючи кожне y_i як показник для елементу i (іншими словами, y_i дорівнює 1, якщо i належить до домену d, або 0 в іншому випадку):

$$\hat{N}_d = \sum_{i \in s_d} \frac{1}{\pi_i} \tag{2.2}$$

Для оцінки середнього значення для усього домену ми використовуємо середнє для вибірки з домену:

$$\tilde{y}_{s_d} = \frac{\sum_{i \in s_d} \frac{y_i}{\pi_i}}{\sum_{i \in s_d} \frac{1}{\pi_i}} = \frac{\hat{t}_{\pi yd}}{\hat{N}_d} \tag{2.3}$$

Як і для оціночної функції Горвіца-Томпсона, тут немає необхідності знати розмір домену. Фактично, навіть якщо цей розмір відомий, ця функція краща за формулу, яка використовує реальний (замість оціночного) розмір домену. Це пов'язане з тим, що є позитивна кореляція між оцінкою підсумкового показника домену та оціночним розміром домену. Це дає взаємну компенсацію у функції 2.3; якщо домен недостатньо представлений у вибірці, то зазвичай заниженими будуть і оцінка підсумку, і оцінка розміру домену. Така компенсація не відбудеться, якщо використовується реальний розмір домену замість оціночного. Тому, оцінюючи підсумок для домену та маючи відомий розмір домену, краще скористатися не оцінкою Горвіца-Томпсона, а добутком середнього значення для домену та реального розміру домену:

$$\hat{t}_{alt,yd} = N_d \tilde{y}_{s_d} = \frac{N_d}{\hat{N}_d} \hat{t}_{\pi yd} \tag{2.4}$$

Дисперсію оцінки 2.4 можна наближено оцінити так:

$$\mathrm{Var}\left(\hat{t}_{alt,yd}\right) \approx \sum_{i \in U_d} \sum_{j \in U_d} \left(\pi_{ij} - \pi_i \pi_j\right) \left(\frac{y_i - \bar{y}_{U_d}}{\pi_i}\right) \left(\frac{y_j - \bar{y}_{U_d}}{\pi_j}\right) \tag{2.5}$$

і переписати цю оцінку у такому вигляді:

$$\hat{\mathrm{Var}}\left(\hat{t}_{alt,yd}\right) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{i \in s_d} \sum_{j \in s_d} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i - \tilde{y}_{s_d}}{\pi_i}\right) \left(\frac{y_j - \tilde{y}_{s_d}}{\pi_j}\right) \tag{2.6}$$

3. Оцінка за допомогою моделі

3.1 Вступ

Інколи обстеження може базуватися на даних, які не становлять безпосереднього інтересу для обстеження, проте можуть певним чином корелювати з досліджуваною змінною. Типовий приклад – *допоміжні дані*, тобто дані, наявні для кожного елмента сукупності. З іншого боку, можуть бути наявні певні *контрольні дані,* тобто для якоїсь змінної, яка сама не є предметом інтересу для обстеження, відоме, наприклад, середнє або підсумкове значення. Якщо є певна кореляція між досліджуваною змінною та допоміжним чи контрольним показником, ми можемо скористатися цим для покращення нашої оцінки. Це можна зробити за допомогою моделі, проте при цьому слід чітко визначити мету такої моделі. Särndal *et al.* (1992, с. 227) добре описують такий підхід:

> "Роль моделі ξ – описати точковий розкид кінцевої сукупності. Ми сподіваємося, що модель ξ достатньо точно описує сукупність. Ми вважаємо, що кінцева сукупність виглядає так, ніби вона була згенерована за моделлю ξ. Однак у жодному разі не робиться припущення, що сукупність насправді була згенерована за моделлю ξ."

Останнє речення може здатися інтуїтивно дивним та протирічним до попереднього речення. Причина в тому, що, як зазначено на початку розділу 1.1, ми виходимо з принципу, що кожний елемент сукупності має фіксоване значення досліджуваної змінної. Оцінка розраховується за допомогою моделі, але не генерується нею, тому ми називаємо її оцінкою *за допомогою моделі*.

3.2 Пропорційна оцінка (ratio estimation)

Дуже часто використовується пропорційна модель співвідношення між досліджуваною змінною *y* та допоміжною змінною *x*:

$$y_i = \beta x_i + e_i \quad \text{where } \operatorname{Var}(e_i) \propto x_i \tag{3.1}$$

При ПВВ рівняння 3.1 набуває вигляд класичної пропорційної моделі:

$$\hat{t}_{Ry} = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} \sum_{i \in U} x_i \tag{3.2}$$

При цьому береться пропорційне співвідношення між досліджуваною змінною та допоміжною змінною у вибірці, яке застосовується до підсумку допоміжної змінної. При ВНІ це співвідношення узагальнюється до пропорційної π-зваженої оцінки Хайека:

$$\hat{t}_{\pi Ry} = \frac{\sum_{i\in s} y_i / \pi_i}{\sum_{i\in s} x_i / \pi_i} \sum_{i\in U} x_i = \hat{t}_{\pi y} \frac{t_x}{\hat{t}_{\pi x}} \tag{3.3}$$

Тепер зрозуміло, що це проста форма калібрації; оцінка обстежуваної змінної множиться на *модельну вагу,* яка є відношенням реального підсумку допоміжної змінної до оцінки цього підсумку. Якщо модельна вага менше 0.5 або більше 2.0, це показує, що вибірка дуже незбалансована відносно допоміжних даних.

Комбінована пропорційна оціночна функція – спеціальний випадок пропорційної π-зваженої оцінки Хайека для СПВВ, який припускає, що пропорційна залежність має однаковий нахил у кількох стратах.

Пропорційна функція не є незміщеною відносно дизайну. Проте для великих вибірок вона є приблизно незміщеною відносно дизайну. Для того, щоб зміщення можно було ігнорувати, рекомендується вибірка не менше 20 елементів.

Дисперсію пропорційної π-зваженої оцінки Хайека можна наближено оцінити так:

$$\mathrm{Var}\left(\hat{t}_{\pi Ry}\right) \approx \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right)\left(y_i - \beta x_i\right)\left(y_j - \beta x_j\right) \tag{3.4}$$

Цю рівність можна переписати так:

$$\hat{\mathrm{Var}}\left(\hat{t}_{\pi Ry}\right) = \left(\frac{t_x}{\hat{t}_{\pi x}}\right)^2 \sum_{i\in s}\sum_{j\in s}\left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\right)\left(y_i - \hat{\beta} x_i\right)\left(y_j - \hat{\beta} x_j\right)$$

$$\text{noting that} \quad \hat{\beta} = \frac{\sum_{i\in s} y_i / \pi_i}{\sum_{i\in s} x_i / \pi_i} \tag{3.5}$$

3.3 Оцінка на основі лінійної регресії

Нехай ми вважаємо, що між досліджуваною та допоміжною змінними існує лінійна залежність, що необов'язково перетинає початок координат:

$$y_i = \alpha + \beta x_i + e_i \quad \text{where } \mathrm{Var}(e_i) = \sigma^2 \ \forall i \tag{3.6}$$

У випадку ПВВ рівність 3.6 дає нам оціночну функцію на основі лінійної регресії:

$$\hat{t}_{Ly} = N\left(\bar{y}_s + \hat{\beta}\left(\bar{x}_U - \bar{x}_s\right)\right) \tag{3.7}$$

При ВНІ це співвідношення узагальнюється до лінійно-регресійної π-зваженої оцінки:

$$\begin{aligned} &\hat{t}_{\pi Ly} = \hat{t}_{\pi y} + \left(t_x - \hat{t}_{\pi x}\right)\hat{\beta} \\ &\text{where } \hat{\beta} = \frac{\sum_{i\in s}\left(x_i - \tilde{x}_s\right)\left(y_i - \tilde{y}_s\right)/\pi_i}{\sum_{i\in s}\left(x_i - \tilde{x}_s\right)^2/\pi_i} \end{aligned} \tag{3.8}$$

Модельна вага для елементу генеральної сукупності *i* є наступною:

$$g_i = \left(\frac{N}{\hat{N}}\right)\left(1 + \frac{\hat{N}\left(\bar{x}_U - \tilde{x}_s\right)\left(x_i - \tilde{x}_s\right)}{\sum_{j\in s}\left(x_j - \tilde{x}_s\right)^2/\pi_j}\right) \tag{3.9}$$

Зауважимо, що модельна вага є різною для кожного елементу генеральної сукупності. Для обстежень домогосподарств модельні ваги зазвичай знаходяться в інтервалі між 0.7 та 1.4, для обстежень підприємств - між 0.3 та 3.0, причому ширший інтервал пояснюється різноманітністю підприємств. Деякі модельні ваги можуть бути негативними. В такому випадку вибірка є незбалансованою, а елемент сукупності різко відхиляється за своїм значенням, причому напрямок відхилу та незбалансованості є однаковим.

Дисперсію лінійно-регресійної π-зваженої оцінки можна наближено оцінити так:

$$\mathrm{Var}\left(\hat{t}_{\pi Ly}\right) \approx \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right)\left(y_i - \alpha - \beta x_i\right)\left(y_j - \alpha - \beta x_j\right) \tag{3.10}$$

Цю рівність можна переписати так:

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{t}_{\pi Ly}\right)=\sum_{i\in s}\sum_{j\in s}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\right)g_i g_j\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)\left(y_j-\hat{\alpha}-\hat{\beta}x_j\right) \tag{3.11}$$

noting that $\hat{\alpha}=\tilde{y}_s-\hat{\beta}\tilde{x}_s$

3.4 Загальна регресійна оцінка з багатьма змінними (GREG)

Нехай ми вважаємо, що існує мультиваріативна лінійна залежність між досліджуваною змінною y_i та вектором допоміжних змінних $\mathbf{x}_i$:

$$y_i=\mathbf{x}_i'\mathbf{B}+e_i \quad \text{where } \mathrm{Var}\left(e_i\right)=\sigma_i^2 \tag{3.12}$$

У випадку ВНІ рівняння 3.12 дає нам загальну регресійну оціночну функцію з багатьма змінними, відому також як узагальнена регресійна оціночна функція (GREG):

$$\begin{aligned}&\hat{t}_{\pi My}=\hat{t}_{\pi y}+\left(\mathbf{t}_{\mathbf{x}}-\hat{\mathbf{t}}_{\pi x}\right)'\hat{\mathbf{B}}\\&\text{where}\quad \hat{\mathbf{B}}=\left(\sum_{i\in s}\frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma_i^2\pi_i}\right)^{-1}\sum_{i\in s}\frac{\mathbf{x}_i y_i}{\sigma_i^2\pi_i}\end{aligned} \tag{3.13}$$

Модельна вага для елементу генеральної сукупності *i* є наступною:

$$g_i=1+\left(\mathbf{t}_x-\hat{\mathbf{t}}_{\pi x}\right)'\left(\sum_{j\in s}\frac{\mathbf{x}_j\mathbf{x}_j'}{\sigma_j^2\pi_j}\right)^{-1}\frac{\mathbf{x}_i}{\sigma_i^2} \tag{3.14}$$

Дисперсію цієї оцінки можна наближено оцінити так::

$$\mathrm{Var}\left(\hat{t}_{\pi My}\right)\approx\sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\left(y_i-\mathbf{x}_i'\mathbf{B}\right)\left(y_j-\mathbf{x}_j'\mathbf{B}\right) \tag{3.15}$$

Цю рівність можна переписати так:

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{t}_{\pi My}\right)=\sum_{i\in s}\sum_{j\in s}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\right)g_i g_j\left(y_i-\mathbf{x}_i'\hat{\mathbf{B}}\right)\left(y_j-\mathbf{x}_j'\hat{\mathbf{B}}\right) \tag{3.16}$$

4. Калібрація

4.1 Оціночна функція GREG як приклад калібрації

Оцінка з використанням калібрації базується на двох принципах. Перший-модельні ваги мають бути якомога ближчими до 1.0 (та зокрема бути позитивними), тобто добуток вагів, на які зважується елемент сукупності, має

наближатися до ваги дизайну цього елементу. Другий принцип полягає в тому, що зважена сума допоміжних векторів має дорівнювати підсумку допоміжних векторів для всієї сукупності. Ця вимога визначає обмеження на калібрацію, яке можна виписати в алгебраїчному виразі:

$$\sum_{i\in s}\frac{g_i}{\pi_i}\mathbf{x}_i=\sum_{i\in U}\mathbf{x}_i \tag{4.1}$$

Пояснення необхідності таких калібраційних обмежень наводиться у Deville & Särndal (1992):

> "калібровані ваги мають давати найкращі оцінки, коли вони застосовуються до кожної допоміжної змінної. Це є перевірка на узгодженість, який цінується багатьма практиками, оскільки сильна кореляція між досліджуваною та допоміжними змінними означає, що ваги, які добре працюють для допоміжних змінних, мають також добре працювати для досліджуваної змінної."

Згадані два принципи комбінуються шляхом мінімізації функції відстані відносно модельної ваги за калібраційних обмежень 4.1. Зазвичай використовується така функція відстані:

$$\sum_{i\in s}\frac{\left(g_i/\pi_i-1/\pi_i\right)^2\sigma_i^2}{1/\pi_i} \tag{4.2}$$

Використання функції відстані 4.2 у поєднанні з калібраційними обмеженнями 4.1 дає нам унікальний набір модельних вагів, ідентичних до тих, що мають місце для оціночної функції GREG (рівняння 3.14). Тому хоча явним чином не використовується жодна модель, проте калібрація з використанням функції відстані 4.2 неявно спирається на загальну модель регресії з багатьма змінними 3.12.

4.2 Калібраційні оціночні функції

У розділі 4.1 ми пояснили, чому функція GREG є прикладом калібраційної оціночної функції. У загальному випадку для калібрації можна використовувати функції відстані, відмінні від 4.2. Необов'язково вказувати явну модель (таку як 3.12), хоча задаючи належні умови регулярності, ми неявно її будуємо (див. Deville & Särndal, 1992 щодо деталізації таких умов та прикладів функції

відстані). Калібрацію можна далі вдосконалювати, встановлюючи додаткові нижню або/та грані на добуток вагів:

$$l_i \le \frac{g_i}{\pi_i} \le u_i \qquad (4.3)$$

Ці грані можуть відрізнятися для різних елементів сукупності. Проте може трапитися, що немає розв'язку, який мінімізував би конкретну функцію відстані за обмежень 4.1 та в діапазоні 4.3.

4.3 Прочісування (ітеративне пропорційне масштабування)

Прочісування (відоме також як "ітеративне пропорційне масштабування") – метод оцінки величин у клітинах таблиці (можливо, багатовимірної), що приводяться у відповідність зі заздалегідь заданими сумами по рядках та колонках; задаються також початкові значення у клітинках. Використовується наступний двомірний алгоритм:

1. Кожна клітинка у першому рядку таблиці помножається на відношення заданого підсумку для цього рядка до суми поточних значень. Таким чином сума значень рядка тепер дорівнюватиме заданій величині.
2. Решта клітинок таблиці масштабується аналогічним чином до заданих підсумків рядків.
3. Усі клітинки масштабуються аналогічним чином за колонками відповідно до заданих підсумків колонок.
4. Етапи 1-3 повторюються до сходження процесу.

Багатовимірний алгоритм працює аналогічним чином – послідовно "прочісується" кожний вимір до досягнення сходження. Слід зазначити, що остаточний розв'язок залежить від початкових значень. Проте прочісування вважається добрим методом, якщо у клітинках немає нульових початкових значень.

Замість описаного алгоритма, можна здійснити прочісування з використанням калібраційної оцінки.

5. Невідповіді та групи однорідності відповідей

У практиці обстежень, як правило, мають місце невідповіді. Іншими словами, щодо певних елементів сукупності, відібраних до вибірки, дані відсутні. Деякі обстеження вважають, що усі елементи мають однакову ймовірність надання відповіді за умови відбору елементу. У такому разі набір відповідей (тобто елементів, щодо яких ми отримали дані) подібний до двоступеневої вибірки, для якої на першому етапі відбираються елементи до вибірки, а на другому отримуються відповіді. Ймовірності включення на другому етапі є рівними для всіх елементів.

У дійсності можна вважати, що ймовірність відповіді корелюється з обстежуваною змінною *y*. Наприклад, в обстеженні закордонних поїздок можна очікувати, що люди менш охоче надаватимуть відповідь у разі візиту до певних країн. Це призведе до суттєвого зміщення оцінки, якщо ми вважаємо що ймовірності відповіді є однаковими. Тому необхідна певна модель для невідповідей. Проблема полягає у недостатності даних про досліджувану змінну від елементів, що не надали відповіді, на яких слід будувати модель. Однак є поста, але корисна модель – групи однорідності відповідей (RHG). За цією моделлю вибірка розбивається на кілька груп. Вважається, що елементи у кожній групі мають однакову ймовірність відповіді, але у різних групах ці ймовірності можуть відрізнятися. Знов-таки, це подібне до двоступеневої вибірки, у якій первинні елементи вибірки – це групи, а вторинні – це власне елементи сукупності, хоча у моделі RHG групи можна будувати і після побудови вибірки. Тому нам необхідно оцінити ймовірність відповіді θ для кожної групи z:

$$\theta_z = \frac{\text{кількість відповідей у групі } z}{\text{кількість відібраних елементів у групі } z} \qquad (5.1)$$

На основі цього можна оцінити сукупну ймовірність π^* для всіх елементів, що надали відповідь:

$$\pi_i^* = \pi_i \theta_z \qquad (5.2)$$

Нарешті, π^* можна підставити замість π у формулах оцінки.

Література

Basu P. (1971) An Essay on the Logical Foundations of Survey Sampling, part one (with discussion), in *Foundations of Statistical Inference* (eds. Godambe, V.P. & Sprott, D.A.). Holt, Rinehart & Winston, Toronto, 203-242.

Cochran, W.G. (1977) *Sampling Techniques.* Wiley, New York.

Deville, J.-C. & Särndal, C.-E. (1992) Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, 376-382.

Horvitz, D.G. & Thompson, D.J. (1952) A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of the American Statistical Association*, **47**, 663-685.

Neyman, J. (1934) On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection, *Journal of the Royal Statistical Society*, **97**, 558-606.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992) *Model Assisted Survey Sampling.* Springer-Verlag, New York.