УДК 519.253 doi: 10.15421/322201

Я.С. Бондаренко
Дніпровський національний університет імені Олеся Гончара

ВИБІРКОВЕ ОБСТЕЖЕННЯ ДОМОГОСПОДАРСТВ STATVILLAGE

В роботі запропоновано методику статистичного аналізу даних домогосподарств гіпотетичного містечка StatVillage. Розглянуто сучасні методи заповнення пропусків в багатовимірних статистичних даних. Порівняно точність оцінок при простому випадковому відборі без повернення, відборі Бернуллі, систематичному відборі та простому випадковому відборі з поверненням.

Ключові слова: вибіркове обстеження, вибірковий дизайн, статистичні дані з пропусками, оцінка Горвіца-Томпсона, дизайн-ефект.

Ya.S. Bondarenko
Oles Honchar Dnipro National University

STATVILLAGE HOUSEHOLDS SURVEY SAMPLING

Methodology of statistical data analysis of households in hypothetical city StatVillage is presented. Advanced techniques for processing multidimensional missing data are examined. Comparisons of estimates accuracy for simple random sampling without replacement, Bernoulli sampling, systematic sampling and simple random sampling with replacement are carried out.

Hypothetical city StatVillage based on real data provides a number of advantages for learning survey sampling methods. Each student of the academic group chooses his own imputation method to handle missing values in statistical data of StatVillage households. There is no universal method for missing values replacement which would be superior in accuracy to all others. Simple imputation methods such as mode imputation and mean imputation have demonstrated performance comparable to advanced imputation methods such as k nearest neighbors, random forest, linear regression, k means. The choice of the imputation method may depend on the types of features with missing values, on the number of objects with missing values, and on the nature of the missing values. Each problem requires an individual approach for imputation missing values.

Each student of the academic group obtains StatVillage maps with marked households for survey sampling. Comparison of the student household maps of the entire academic group illustrates the differences in sampling designs therefore the variation in the values of the parameter estimates arises from the random selection method of households for the sample.

Each student of the academic group obtains point estimates and interval estimates of parameters for their own StatVillage household survey. Point and interval estimates obtained by students of the entire academic group make it possible to construct histograms of distributions for values of point estimates and a set of confidence intervals which contains unknown parameters with a given probability.

Each student of the academic group obtains his own estimates of the design-effect as a measure of the effectiveness of selection strategies using a variety of designs. Estimates of the design-effect obtained by students of the entire academic group allow choose the optimal sample design for household survey in the hypothetical city StatVillage.

Keywords: survey sampling, sampling design, statistical data with missing values, Horvitz-Tompson estimates, design-effect.

Я.С. Бондаренко
Днипровский национальный университет имени Олеся Гончара

ВЫБОРОЧНОЕ ОБСЛЕДОВАНИЕ ДОМОХОЗЯЙСТВ STATVILLAGE

В работе предложена методика статистического анализа данных домохозяйств гипотетического городка StatVillage. Рассмотрены современные методы заполнения пропусков в многомерных статистических данных. Проведено сравнение точности оценок при простом случайном отборе без возвращения, отборе Бернулли, систематическом отборе и простом случайном отборе с возвращением.

Ключевые слова: выборочное обследование, выборочный дизайн, статистические данные с пропусками, оценка Горвица-Томпсона, дизайн-эффект.

Вступ. Гіпотетичне містечко StatVillage, побудоване на основі реальних статистичних даних, доцільно використовувати при навчанні студентів методам вибіркових обстежень [1]. Студентам надається клікабельна мапа містечка, на якій розташовані домогосподарства. Після того, як студенти позначають житлові будинки для обстеження, вони надсилають запит на сервер, який повертає статистичні дані, які є фактичними спостереженнями домогосподарств, здобутими Статистичним Управлінням Канади підчас перепису 1991 року [2]. Містечко StatVillage має три інтерактивні мапи. Максимальна конфігурація Maximal Village містечка StatVillage складається з масиву зі 128 блоків, причому кожен блок складається з 8 житлових будинків, загалом 1024 будинка. У цій конфігурації використовується простий макет для того, щоб студенти зосередилися на порівнянні вибіркових дизайнів в простих популяціях, перш ніж переходити до більш складних популяцій. Тим не менш, такий макет дозволяє студентам відбирати одиниці спостережень за допомогою простого випадкового відбору, систематичного відбору, кластерного відбору, двостадійного і двофазного відборів. Мінімальна конфігурація Mini Village містечка StatVillage складається з 60 блоків, кожен блок містить 8 житлових будинків, загалом 480 будинків. Мікро конфігурація Micro Village містечка StatVillage складається з 36 блоків, кожен блок містить 8 житлових будинків, загалом 288 будинків. Mini Village та Micro Village корисні в ситуаціях, коли проблеми з технічним або програмним забезпеченням не дозволяють працювати з максимальною конфігурацією містечка StatVillage.

Статистичні дані для побудови містечка StatVillage були відібрані навмання із записів домогосподарств (односімейні будинки або одноповерхові квартири у Ванкувері, Британській Колумбії, Канаді) з файлу мікроданих перепису населення 1991 року для громадського користування про домогосподарства та житло [2]. Файл мікроданих містить анонімні відповіді, зокрема, демографічні змінні – розмір домогосподарства та склад домогосподарства за віковими показниками та статтю; змінні доходу – зайнятість, інвестиції, державні трансферти тощо; характеристики житла – тип, рік побудови, власне чи

орендоване, орієнтовна вартість, місячна вартість проживання тощо; характеристики що найменше двох утримувачів домогосподарства (дорослі, відповідальні за добробут домогосподарства) – вік, стать, рід занять, рідна мова, освіта, статус зайнятості тощо.

Навчальна програма курсу з методів вибіркових обстежень складається з низки теоретичних відомостей, зокрема, основних формул для оцінок параметрів популяції, дисперсій цих оцінок та оцінок дисперсій, та практичних задач, при цьому витрати часу на фактичний збір даних або розробку дизайнів обстежень невеликі. Методиці збору даних для простих вибіркових обстежень присвячена низка джерел [3-12], але важко змусити студентів збирати дані за більш складними вибірковими схемами. Ще важче повторити збір даних щодо однієї реальної популяції, використовуючи різні вибіркові схеми для того, щоб студенти могли порівняти дизайни на практиці. Тому для усвідомлення студентами особливостей, які виникають при аналізі реальних статистичних даних, наполегливо рекомендується використовувати гіпотетичне містечко StatVillage при вивченні методів вибіркових обстежень.

Постановка задачі. Розробити методику проведення статистичного аналізу даних домогосподарств гіпотетичного містечка StatVillage.

Обсяг вибірки. Визначення обсягу вибірки є ітераційною процедурою, складність якої полягає у тому, що практично всі параметри оцінюються або експертно, або за даними пілотних обстежень. Методика формування вибірки для проведення вибіркових обстежень населення складається з таких етапів:

1) обчислюються $\hat{\bar{y}}$ – оцінка середнього значення ознаки, $\hat{Y}$ – оцінка сумарного значення ознаки, $\hat{\omega}$ – оцінка частки елементів з певною властивістю;

2) оцінюється значення дисперсії σ^2_{srs} кожної ознаки за умови побудови вибірки за допомогою простого випадкового відбору;

3) проводиться оцінка дизайн-ефекту $\hat{deff}$ (відношення варіації оцінки ознаки для реального дизайну вибірки до варіації оцінки цієї ж ознаки за умови побудови вибірки за допомогою простого випадкового відбору);

4) встановлюються вимоги щодо надійності оцінок ознак (коефіцієнт варіації CV оцінки обирається від 1% до 10%);

5) розраховуються обсяги вибірки для отримання надійних оцінок для кожної ознаки.

Обсяг вибірки для оцінювання середнього значення дорівнює

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

Обсяг вибірки для оцінювання сумарного значення дорівнює

$$n = \hat{deff} \cdot N^2 \cdot \hat{\sigma}^2_{srs} \cdot \left(\frac{100\%}{CV \cdot \hat{Y}} \right)^2 .$$

Обсяг вибірки для оцінювання частки елементів з певною властивістю дорівнює

$$n = \hat{deff} \cdot \hat{\sigma}^2_{srs} \cdot \left(\frac{100\%}{CV \cdot \hat{\omega}} \right)^2 .$$

6) із розрахованих обсягів вибирається максимальний;
7) визначається прийнятний обсяг вибірки з урахуванням організаційних та фінансових можливостей. У випадку недостатності коштів для забезпечення проведення обстеження необхідно зменшити надійність оцінок показників (пункт 4), після цього розрахувати нові обсяги вибірки (пункт 5) і повторити пункти 6, 7;
8) проводиться стратифікація вибірки за умови наявності страт;
9) проводиться коригування обсягу вибірки з урахування навантаження інтерв'юера;
10) проводиться коригування обсягу вибірки у зв'язку з відмовами респондентів.

Наведена схема обчислення обсягу вибірки виявила свою ефективність при формуванні вибірок для проведення вибіркових обстежень населення [13].

Методи заповнення пропусків в багатовимірних даних. Нехай статистичні дані представлені матрицею об'єктів-ознак $X^{(n\times d)}$, де n – кількість об'єктів, d – кількість ознак. Частина значень матриці ознак відсутні. Необхідно відновити матрицю об'єктів-ознак з метою подальшого застосування методів вибіркових обстежень.

Найпростішим методом розв'язання задачі обробки пропущених значень є видалення об'єктів (рядків матриці об'єктів-ознак), що мають пропуски в даних. Метод застосовується лише у тому випадку, коли невелика частина об'єктів вибірки має пропущені значення. Недоліками даного методу можна вважати втрату інформації при виключенні об'єктів. Видалення ознак (стовбців матриці об'єктів-ознак), що мають пропуски в даних, є альтернативним методом при наявних пропусках у невеликій кількості ознак.

Другим найпростішим методом заповнення пропусків є заповнення пропущених даних *модою* або *середнім значенням* за кожною ознакою. Як правило, пропуски в категоріальних ознаках заповнюють модою, в кількісних ознаках – середнім значенням [15].

Третім найпростішим методом заповнення пропущених даних є заповнення пропусків за допомогою *сингулярного розкладу матриці ознак*. Спочатку відбувається заміна пропусків вибірковим середнім значенням, обчисленим за даними без пропусків за кожною ознакою матриці об'єктів-ознак $X^{(n\times d)}$ і для збереження природності даних значення в заповнених пропусках замінюються на найближчі унікальні значення за кожною ознакою.

Знаходиться розкладання матриці ознак $X^{(n\times d)}$ у вигляді

$$X^{(n\times d)} = U^{(n\times n)} S^{(n\times d)} V^{(d\times d)},$$

де $S^{(n\times d)}$ – сингулярна матриця, тобто діагональна матриця, на головній діагоналі якої розташовані корені з власних значень матриці $X^{(n\times d)}\left(X^{(n\times d)}\right)^T$ в

порядку спадання. Матриці $U^{(n\times n)}$, $V^{(d\times d)}$ є ортогональними, при цьому стовбці матриці $U^{(n\times n)}$ є власними векторами матриці $X^{(n\times d)}$, а матрицю $V^{(d\times d)}$ можна подати у вигляді $V^{(d\times d)}=\left(S^{(n\times d)}\right)^{-1}\left(U^{(n\times n)}\right)^{T}X^{(n\times d)}$.

Перші r рядків та стовбців виділяються в матриці $S^{(n\times d)}$, а ті, що залишилися видаляються. Перші r значущих сингулярних чисел називаються головними компонентами.

Виділивши в матриці $U^{(n\times n)}$ перші r стовбців – $U^{(n\times r)}$, а в матриці $V^{(d\times d)}$ перші r рядків – $V^{(r\times d)}$, можна відновити матрицю ознак

$$X_{approx}^{(n\times d)}=U^{(n\times r)}S^{(r\times r)}V^{(r\times d)}.$$

Відбувається заміна значень в заповнених пропусках в матриці ознак $X^{(n\times d)}$ на значення, здобуті у відновленій матриці $X_{approx}^{(n\times d)}$.

Кроки 2, 3, 4 можна повторити заздалегідь задане число разів для покращення відновлення матриці ознак або скористатися критерієм якості відновлення матриці, обчисливши близькість до одиниці коефіцієнта детермінації:

$$Q(r)=\frac{\sum_{k=1}^{r}\lambda_k}{\sum_{k=1}^{n}\lambda_k},$$

де λ_k – власні значення матриці $X^{(n\times d)}\left(X^{(n\times d)}\right)^{T}$. Залежність коефіцієнта детермінації $Q(r)$ від числа головних компонент r дозволяє оцінити ефективність методу.

Наприкінці значення в заповнених пропусках замінюються на найближчі унікальні значення за кожною ознакою матриці $X_{approx}^{(n\times d)}$ для збереження природності даних [15].

При заповненні пропусків за допомогою *методу k найближчих сусідів* висувається гіпотеза про те, що близькі об’єкти мають близькі значення ознак. Тому пропущені значення ознак певного об’єкту можна відновити за відомими значеннями ознак k «найближчих сусідів» цього об’єкту [15].

При заповненні пропусків за допомогою *випадкового лісу* для кожної ознаки з пропусками розв’язується задача прогнозування за допомогою випадкового лісу (при цьому навчання проводиться по об’єктах без пропусків по даній ознаці). Заміна пропущених значень в кожній ознаці проводиться за допомогою прогнозу композиції дерев, здобутому при розв’язанні задачі прогнозування [15].

При заповненні пропусків за допомогою *лінійної регресії* для кожної ознаки з пропусками розв’язується задача прогнозування за допомогою лінійної регресії (при цьому навчання проводиться по об’єктах без пропусків по даній ознаці). Заміна пропущених значень в кожній ознаці проводиться за допомогою прогнозу, здобутому при розв’язанні задачі лінійної регресії [15].

При заповненні пропусків за допомогою *методу* k *середніх* висувається гіпотеза про те, що близькі об'єкти мають близькі значення ознак. Тому пропущені значення ознак певного об'єкту можна відновити за відомими значеннями ознак центру кластеру, якому належить об'єкт з пропусками [15].

Простий випадковий відбір без повернення. Обсяг вибірки для ПВВбП обчислюється за допомогою методики формування вибірок. Оцінки середнього значення ознаки $\hat{\bar{y}}$, сумарного значення ознаки $\hat{Y}$, частки одиниць з певною властивістю $\hat{\omega}$, дисперсії $\hat{\sigma}^2_{srs}$ знаходяться за статистичними даними містечка Micro Village. Оцінка дизайн-ефекту та коефіцієнту варіації прийняті $\hat{deff} = 1$, $CV = 5\%$ відповідно. Обсяг вибірки становить n_1. Проста випадкова вибірка обсягом n_1 домогосподарств відбирається з містечка Maximal Village. Знаходяться оцінки Горвіца-Томпсона $\hat{t}_\pi$, $\hat{\bar{y}}_\pi$, $\hat{P}_d$ при ПВВбП для сумарного значення, середнього значення та частки елементів з певною властивістю [9].

Відбір Бернуллі. Обсяг вибірки для відбору Бернуллі обчислюється за допомогою методики формування вибірок. Оцінки середнього значення ознаки $\hat{\bar{y}}$, сумарного значення ознаки $\hat{Y}$, частки одиниць з певною властивістю $\hat{\omega}$, дисперсії $\hat{\sigma}^2_{srs}$, дизайн-ефекту $\hat{deff}$ знаходяться за статистичними даними містечка Micro Village. Коефіцієнт варіації CV дорівнює 5%. Обсяг вибірки становить n_2. Вибірка обсягом n_2 домогосподарств відбирається з містечка Maximal Village за допомогою відбору Бернуллі (імовірність включення домогосподарства у вибірку становить 0,5). Знаходяться оцінки Горвіца-Томпсона $\hat{t}_\pi$, $\hat{\bar{y}}_\pi$, $\hat{P}_d$ при відборі Бернуллі для сумарного значення, середнього значення та частки елементів з певною властивістю [9].

Систематичний відбір. Обсяг вибірки для систематичного відбору обчислюється за допомогою методики формування вибірок. Оцінки середнього значення ознаки $\hat{\bar{y}}$, сумарного значення ознаки $\hat{Y}$, частки одиниць з певною властивістю $\hat{\omega}$, дисперсії $\hat{\sigma}^2_{srs}$, дизайн-ефекту $\hat{deff}$ знаходяться за статистичними даними містечка Micro Village. Коефіцієнт варіації CV дорівнює 5%. Обсяг вибірки становить n_3. Вибірка обсягом n_3 домогосподарств відбирається з містечка Maximal Village за допомогою систематичного відбору. Знаходяться оцінки Горвіца-Томпсона при систематичному відборі для сумарного значення, середнього значення та частки елементів з певною властивістю [9].

Простий випадковий відбір з поверненням. Обсяг вибірки для ПВВзП обчислюється за допомогою методики формування вибірок. Оцінки середнього значення ознаки $\hat{\bar{y}}$, сумарного значення ознаки $\hat{Y}$, частки одиниць з певною властивістю $\hat{\omega}$, дисперсії $\hat{\sigma}^2_{srs}$, дизайн-ефекту $\hat{deff}$ знаходяться за статистичними даними містечка Micro Village. Обсяг вибірки для простого випадкового відбору з поверненням становить n_4. Проста випадкова вибірка

без повернення відбирається з містечка Maximal Village за допомогою метода накопичених сум. Знаходяться оцінки Горвіца-Томпсона $\hat{t}_\pi$, $\hat{\bar{y}}_\pi$, $\hat{P}_d$ для сумарного значення, середнього значення та частки елементів з певною властивістю [9].

Оцінка дизайн-ефекту. У теорії вибіркових обстежень простому випадковому відбору без повернення надана роль еталона, з яким порівнюють усі інші відбори. Обчислимо оцінку дизайн ефекту:

$$deff\,(p(s),\hat{t}_\pi) = \frac{\hat{D}\hat{t}_\pi}{\hat{D}_{ПВВбП}\hat{t}_\pi},$$

де $\hat{D}\hat{t}_\pi$ – оцінка дисперсії відбору за допомогою вибіркового дизайну $p(s)$, $\hat{D}_{ПВВбП}\hat{t}_\pi$ – оцінка дисперсії при ПВВбП. Дизайн-ефект визначає ефективність стратегії відбору за допомогою вибіркового дизайну $p(s)$ порівняно з ПВВбП. А саме: якщо

$$deff\,(p(s),\hat{t}_\pi) < 1,$$

то здобуваємо точніші результати за допомогою вибіркового дизайну $p(s)$ порівняно з ПВВбП, в супротивному разі – ні. Оскільки в обох випадках використовується оцінка Горвіца-Томпсона $\hat{t}_\pi$, то причиною відмінностей дисперсій оцінок є вибірковий дизайн $p(s)$. Аналогічні міркування мають місце для оцінок $\hat{\bar{y}}_\pi$ та $\hat{P}_d$.

Статистичний аналіз даних гіпотетичного містечка StatVillage виконано за допомогою мови програмування R. Здобуті результати говорять про ефективність систематичного відбору при оцінюванні таких ознак домогосподарств StatVillage як вартість житла, вартість комунальних послуг за місяць, загальний дохід, кількість жителів домогосподарств та наявність іпотеки.

Висновки. Гіпотетичне містечко StatVillage надає низку переваг для вивчення курсу з методів вибіркових обстеженень.

Кожен студент академічної групи обирає власний метод заповнення пропусків в статистичних даних домогосподарств містечка StatVillage. Не існує універсального методу заповнення пропущенних значень, який би перевершував за якістю інші методи. Вибір методу може залежати від типів ознак, у яких виникають пропуски, від кількості об'єктів з пропущеними значеннями, від природи виникнення пропусків. У кожній задачі необхідний індивідуальний підхід до вибору методу відновлення пропущених значень.

Кожен студент академічної групи здобуває мапи містечка StatVillage з позначеними домогосподарствами для обстеження. Порівняння мап домогосподарств студентів всієї академічної групи ілюструє відмінності вибіркових дизайнів, отже, варіація значень оцінок параметрів виникає внаслідок випадкового методу відбору домогосподарств до вибірки.

Кожен студент академічної групи здобуває точкові та інтервальні оцінки параметрів власного обстеження домогосподарств StatVillage. Точкові та інтервальні оцінки параметрів, здобуті студентами всієї академічної групи, до-

зволяють побудувати гістограми розподілів значень точкових оцінок параметрів та множини довірчих інтервалів, які містять невідомі параметри з заданою ймовірністю.

Кожен студент академічної групи здобуває власні оцінки дизайн-ефектів як міри ефективності стратегій відборів за допомогою різноманітних дизайнів. Оцінки дизайн-ефектів, здобуті студентами всієї академічної групи, дозволяють обрати оптимальний вибірковий дизайн вибіркового обстеження домогосподартсв гіпотетичного містечка StatVillage.

Бібліографічні посилання

1. Schwarz, C. J. StatVillage: An On-Line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling. *Journal of Statistics Education 5(2), 1997*.
2. Statistics Canada, User Documentation for Public Use Microdata File on Households and Housing, Ottawa, Ontario: Statistics Canada, 1994.
3. J. G. Bethlehem. Applied survey methods: a statistical perspective, John Wiley & Sons, New York, 2009.
4. W. G. Cochran. Sampling Techniques, John Wiley & Sons, New York, 1977.
5. P.S. Levy, S. Lemeshov. Sampling of Population: Methods and Applications, 4th Edition, John Wiley & Sons, New York, 2008.
6. S.L. Lohr. Sampling: Design and Analysis, Cengage Learning, 2009.
7. C.-E. Sarndal, B. Swensson, J. Wretman. Model Assisted Survey Sampling, Springer, New York, 1992.
8. R.L. Scheaffer, III W. Mendenhall, R.L. Ott, K. G. Gerow. Elementary Survey Sampling, 7th Edition, Cengage Learning, 2011.
9. Василик О.І., Яковенко Т.О. Лекції з теорії і методів вибіркових обстежень: навчальний посібник. К.: Видавничо-поліграфічний центр «Київський університет», 2010. 208 с.
10. Пархоменко В.М. Методи вибіркових обстежень. К.:ТВіМС, 2001. 148 с.
11. Черняк О.І. Техніка вибіркових обстежен. К.: МІВВЦ, 2001. 248 с.
12. Саріогло В.Г. Проблеми статистичного зважування вибіркових даних. К.: ІВЦ Держкомстату України, 2005. 264 с.
13. Гладун О.М. Вибіркові обстеження населення: методологія, методика, практика: Монографія. Ніжин: ТОВ Видавництво Аспект-Поліграф, 2008. 348 с.
14. Little, R. J. A., Rubin, D. B. Statistical Analysis with Missing Data. John Wiley & Sons, New York, 2002.
15. Bondarenko Yana Statistical Analysis with Missing Data. *Proceedings of the Baltic-Nordic-Ukrainian Workshop on Survey Statistics 2022*. Tartu, Estonia, 2022, pp. 35-39.

Надійшла до редколегії 08.09.2022.