

Розділ 5. Проблемні питання прикладних досліджень

5.1. Формування вибірки

5.1.1. Основні методи формування вибірки

Формування вибірки – це базовий етап будь-якого прикладного дослідження (основні вимоги до вибірки викладено у п. 1.1.). При формуванні вибірки досліднику необхідно визначити:

– хто (що) є елементом або *одиноцею вибірки* виходячи від сутності дослідження;

– *контур вибірки*, тобто список усіх одиниць генеральної сукупності, з якої формується вибірка;

– *об'єм вибірки* – кількість елементів у ній.

Наприклад, якщо фірма – виробник мобільних телефонів бажає вивчити потенціальний ринок своєї продукції, то одиницями вибірки будуть особи, які приймають рішення про вибір комунікаційного обладнання: керівники організацій або голови родин. Контуром вибірки можуть бути списки: організацій, фірм, домовласників і т.п.

Оскільки вибірка є лише частиною генеральної сукупності, то отримані на основі її вивчення результати не будуть точно відповідати результатам, які можна було б отримати при вивченні всієї генеральної сукупності. Різниця між результатами дослідження вибірки та генеральної сукупності називається *помилкою вибірки*. Помилки вибірки обумовлюються як методами її формування, так і її об'ємом.

Ймовірнісні методи формування вибірки

При формуванні вибірки часто використовують ймовірнісні методи, при яких всі елементи генеральної сукупності мають однакову ймовірність бути включеними в вибірку. Такими методами є простий випадковий відбір, систематичний відбір, кластерний відбір та стратифікований відбір.

При *простому випадковому відборі* передбачається, що для всіх елементів генеральної сукупності ймовірність бути обраним в вибірку відома і однакова. Така ймовірність визначається як відношення об'єму вибірки до об'єму генеральної сукупності.

Простий випадковий відбір може здійснюватись за допомогою таблиць випадкових чисел або генератора випадкових чисел MS Excel. При цьому: кожному елементу генеральної сукупності присвоюється порядковий номер; генерується необхідна кількість випадкових чисел; обираються ті елементи генеральної сукупності, номери яких співпадають з випадковими числами. Так, наприклад, при проведенні телефонного інтерв'ю комп'ютер випадково генерує телефонні номери.

Недоліком цього методу є необхідність визначення кожного елемента генеральної сукупності, що часто просто неможливо.

При *систематичному відборі* необхідно: визначити всі елементи генеральної сукупності; визначити інтервал відбору – відношення об'єму генеральної сукупності до об'єму вибірки; відібрати в вибірку елементи генеральної сукупності з урахуванням інтервалу.

Наприклад, якщо як генеральна сукупність використовується телефонний довідник, в якому 5500 номерів, а необхідний об'єм вибірки дорівнює 500, то інтервал відбору визначається як $5500 / 500 = 11$. Це означає, що кожний одинадцятий телефонний номер потрапляє у вибірку.

Кластерний відбір базується на розподілі елементів генеральної сукупності на групи (кластери), кожна з яких репрезентативна всій сукупності. У подальшому випадково обирається один із кластерів, елементи якого вважаються генеральною сукупністю. На ньому проводиться певне дослідження, результати якого розповсюджуються на всі інші кластери і на всю генеральну сукупність.

Наприклад, якщо планується дослідження думки населення певної області, то кластерами можуть вважатися райони цієї області. При цьому дослідник припускає, що населення кожного району ідентичне населенню області у цілому (що не завжди правильно).

У випадку, коли дослідження проводиться на певній території, можна розбити цю територію на кластери за допомогою вибіркової решітки, яка накладається на карту території і визначає кластери.

Недоліком усіх описаних методів є необхідність припущення про симетричний розподіл характеристик генеральної сукупності, що практично зустрічається дуже рідко. Наприклад, населення різних районів Одеської області значно відрізняється за рівнем доходів, основними видами діяльності і т. ін., тому її райони не можуть вважатися ідентичними як один одному, так і всій області в цілому.

Якщо розподіл характеристик генеральної сукупності не є симетричним, використовується **стратифікований відбір**. В цьому випадку генеральна сукупність розподіляється на групи (страти) залежно від певної характеристики. Наприклад, населення області розподіляється на групи залежно від рівня доходу. Кожна страта виконує роль генеральної сукупності, з якої формується вибірка.

При цьому необхідно враховувати пропорційність стратифікування. Якщо об'єм вибірки для певної страти пропорційний розміру страти у відношенні до всієї генеральної сукупності, то вибірка називається пропорційно стратифікованою. Якщо ця умова не виконується, то слід застосовувати вагові коефіцієнти, які урівноважать розміри страт.

Неймовірнісні методи формування вибірки

Використання ймовірнісних методів формування вибірки є найбільш правильним з точки зору статистики, але вони трудомісткі і, відповідно, матеріально не вигідні. Тому в прикладних дослідженнях часто використовують неймовірнісні методи: відбір за критерієм, відбір згідно спостереження, відбір за квотами та відбір за принципом зручності.

Метод **відбору за принципом зручності** полягає в тому, що формування вибірки здійснюється самим зручним з боку дослідника способом за мінімум часу та грошових витрат. Наприклад, опитування покупців проводиться в певному магазині. При цьому вибірка часто не є репрезентативною, але існують способи оцінки її похибки.

При **формуванні вибірки за критеріями** використовується думка спеціалістів-експертів щодо її складу. Так часто формуються фокус-групи.

Метод **відбору за спостереженням** використовується тоді, коли контури вибірки дуже обмежені. Наприклад, при проведенні дослідження попиту на продукцію, що не має широкого кола споживачів.

За цим методом дослідник формує початкову вибірку, об'єм якої значно менший необхідного. У процесі дослідження вибірка розширюється за рахунок пропозицій і рекомендацій респондентів, які вже прийняли участь в опитуванні.

Метод **відбору за квотами** використовується за умови наявності чітко визначених характеристик респондентів, думку яких доцільно вивчити при дослідженні. У цьому випадку в вибірку відбирають ті елементи генеральної сукупності, що мають певні характеристики. Об'єм вибірки (квота) також визначається на початку дослідження. Відбір елементів закінчується при заповненні квоти.

Зауваження. При формуванні вибірки дослідник повинен знайти баланс між затратами на збір даних і об'ємом вибірки. Методи формування вибірки повинні відповідати цілям дослідження.

5.1.2. Визначення об'єму вибірки

В практичних дослідженнях використовується кілька методів розрахунку об'єму вибірки. Але незалежно від результатів розрахунків слід пам'ятати, що:

- об'єм вибірки залежить передусім від вартості дослідження;
- точність отриманих результатів залежить не стільки від об'єму вибірки, скільки від методу її формування.

Об'єм вибірки може бути просто процентом від об'єму генеральної сукупності. Наприклад, припускається, що для отримання досить точних результатів достатньо 5% елементів генеральної сукупності. Однак перевірити таке припущення неможливо, тому рівень точності результатів невідомий.

Якщо певне дослідження проводиться регулярно, то доцільно використовувати вибірки однакового об'єму. Крім того, якщо відомо, що дослідження певного типу вже проводилося, можна обрати той же об'єм вибірки. Але при цьому не враховуються можливі зміни умов проведення дослідження.

При проведенні соціологічних опитувань вважається, що для пробних (пілотних) опитувань достатня вибірка об'ємом 100 – 250 осіб. При масових опитуваннях (генеральна сукупність становить 5000 осіб і більше) об'єм вибіркової сукупності повинен становити 10% генеральної сукупності, але не більше 2 – 2,5 тис. осіб. Це гарантує достатньо достовірні результати дослідження. Помилки вибірки, які інколи при цьому трапляються, бувають наслідком невірних вихідних статистичних даних про параметри контрольних ознак генеральної сукупності; недостатнього об'єму вибіркової сукупності, неправильного застосування способу відбору одиниць аналізу (наприклад, відбір із неправильно складеного списку, невдалий вибір місця, часу проведення опитування тощо).

Найбільш коректним є статистичний метод розрахунку об'єму вибірки, заснований на визначенні мінімально необхідного об'єму вибірки залежно від вимог до точності результатів дослідження.

У статистичному методі використовуються такі позначення:

n – об'єм вибірки;

N – об'єм генеральної сукупності;

z – нормативне відхилення оцінки. Обирається залежно від рівня значущості (табл. 5.1). Зазвичай потрібен рівень значущості $\alpha = 0,05$, тоді $z = 1,96$;

S^2 – дисперсія вибірки;

S – середнє квадратичне відхилення вибірки;

p – варіація вибірки. Наприклад, якщо відомо, що 70% населення не вживають йогурт, то варіація вибірки $p = 70$. Якщо варіація невідома, то приймається $p = 50$;

$q = 100 - p$;

e – допустима помилка, яка обирається дослідником.

Таблиця 5.1

α	0,4	0,3	0,2	0,15	0,1	0,05	0,03	0,01	0,003
z	0,84	1,03	1,29	1,44	1,65	1,96	2,18	2,58	3,0

Зауваження. Для визначення варіації певної генеральної сукупності доцільно провести попереднє (пілотне) дослідження. Крім того, слід враховувати, що максимальною є варіація $p = 50$. Тобто така варіація відповідає найгіршому випадку.

Наведемо розрахункові формули для визначення об'єму вибірки у випадку, коли вибірка становить менше 5% від генеральної сукупності і вважається великою:

$$n = \frac{z^2 pq}{e^2} . \quad (5.1)$$

Формула (5.1) використовується, якщо потрібно розрахувати кількість респондентів соціологічного опитування за умов, що на питання існує два варіанти відповіді «Так» або «Ні».

$$n = \frac{z^2 S^2}{e^2} . \quad (5.2)$$

Формула (5.2) використовується, якщо з попередніх досліджень відома дисперсія або середнє квадратичне відхилення.

Якщо такі дослідження не проводилися і вибірка формується вперше, то використовується формула, в якій помилка e пов'язана із середнім квадратичним відхиленням:

$$n = \frac{z^2}{e_1^2}, \text{ де } e_1^2 = \frac{e}{S} . \quad (5.3)$$

Якщо об'єм вибірки більше 5% генеральної сукупності, то вибірка вважається маленькою і в формули 5.1 – 5.3 вводиться уточнюючий коефіцієнт:

$$k = \sqrt{\frac{N-n}{N-1}}. \quad (5.4)$$

Тоді об'єм вибірки n_k розраховується за формулою:

$$n_k = n \cdot k = n \cdot \sqrt{\frac{N-n}{N-1}}. \quad (5.5)$$

Приклад 5.1. Необхідно визначити кількість респондентів для опитування при дослідженні ринку копченої риби.

Розв'язок. Якщо кількість споживачів невідома, приймаємо варіацію вибірки $p = 50$. Рівень значущості $\alpha = 0,05$, тоді $z = 1,96$, припустима помилка 4%. Тоді за формулою 5.1 маємо:

$$n = \frac{z^2 pq}{e^2} = \frac{1,96^2 \cdot 50 \cdot 50}{4^2} = 600 \text{ осіб.}$$

Якщо потрібна менша помилка, наприклад 3%, то за тією ж формулою маємо:

$$n = \frac{z^2 pq}{e^2} = \frac{1,96^2 \cdot 50 \cdot 50}{3^2} = 1067 \text{ осіб.}$$

Якщо проводилося попереднє дослідження і відомо, що 70% респондентів є споживачами, то об'єм вибірки дорівнює:

$$n = \frac{z^2 pq}{e^2} = \frac{1,96^2 \cdot 70 \cdot 30}{4^2} = 504 \text{ особи.}$$

Приклад 5.2. Проводиться дослідження якості послуг, які надаються перукарнями міста. Необхідно визначити кількість респондентів для опитування, якщо відомо, що у місті 500 перукарень.

Розв'язок. Приймаємо варіацію вибірки $p = 50$. Рівень значущості $\alpha = 0,05$, тоді $z = 1,96$, припустима помилка 10%. Тоді за формулою 5.1 маємо:

$$n = \frac{z^2 pq}{e^2} = \frac{1,96^2 \cdot 50 \cdot 50}{10^2} = 96 \text{ перукарень.}$$

У даному випадку об'єм вибірки становить 19% від генеральної сукупності (загальна кількість перукарень – 500 од.), тобто перевищує 5%. Тому розрахуємо об'єм вибірки з урахування уточнюючого коефіцієнта за формулою 5.5:

$$n_k = n \cdot k = 96 \cdot \sqrt{\frac{500-96}{500-1}} = 86 \text{ перукарень.}$$