Тема 4. Вибірка та вибіркові розподіли

4.1. Генеральна сукупність та вибірка

У статистиці *генеральною сукупністю* називають загальну групу об'єктів (людей, предметів, організацій), про яку потрібно отримати інформацію на основі дослідження статистичних даних. *Вибіркою* називають невеликий набір об'єктів, відібраних з генеральної сукупності.

Для позначення певної характеристики генеральної сукупності будемо використовувати термін «параметр», для відповідної характеристики вибірки – термін «статистика». Параметр є константою, статистика – випадкова величина, оскільки її розраховують за вибіркою і для різних вибірок вона може набувати різних значень. Прикладами вибіркових статистик є вибіркова середня, медіана, стандартне відхилення вибірки. Вибіркову статистику використовують як оцінку відповідної характеристики генеральної сукупності. Вибірковий метод у статистичних дослідженнях використовується для отримання оцінки параметра генеральної сукупності на основі вибіркової статистики.

Переваги використання вибіркового методу полягають у наступному.

1. Практичність. Генеральна сукупність може бути дуже великою, тому фізично неможливо виміряти всі її об'єкти.
2. Економія часу. Обчислення вибіркової статистики займає набагато менше часу, ніж обчислення параметра генеральної сукупності.
3. Зменшення витрат на статистичне дослідження.
4. Зменшення кількості помилок реєстрації.
5. Деякі статистичні обстеження об'єктів пов'язані зі знищенням зразків (перевірка терміну горіння електричних ламп, перевірка якості продуктів харчування).

Для застосування у статистичному дослідженні вибірка повинна бути *репрезентативною*, тобто кожна властивість у вибірці та генеральній сукупності повинна мати однакові частки. Якщо вибірка не є репрезентативною, то кажуть, що вона має зміщення.

Організація проведення вибіркового спостереження та обробки його результатів – це процес, що складається з наступних етапів.

1. Обгрунтування методу дослідження.
2. Визначення способу проведення спостереження.
3. Встановлення граничної похибки вибірки.
4. Визначення необхідного обсягу вибірки.
5. Формування вибіркової сукупності.
6. Проведення статистичного спостереження.
7. Обчислення точкових оцінок вибірки.

8. Визначення інтервальних оцінок.
9. Перевірка надійності результатів вибіркового спостереження.
10. Висновки про параметри генеральної сукупності.

У процесі вибіркового дослідження розглядають два типи вибіркових оцінок – точкові та інтервальні. *Точкова оцінка* характеризує значення параметра, обчислене на основі даних вибірки, а *інтервальна* – довірчий інтервал (інтервал значень параметра для даної сукупності).

Якщо значення ознаки у статистичній сукупності формується під впливом великої кількості факторів, жоден з яких не є переважаючим у порівняння з іншими факторами, то кажуть, що така сукупність має *нормальний розподіл*.

Якщо генеральна сукупність має нормальний розподіл, то розподіл вибіркових характеристик також буде нормальним. Навіть у випадку невідповідності розподілу генеральної сукупності нормальному розподілу для вибірок великого обсягу ($n \geq 30$) розподіл вибіркових середніх приблизно відповідатиме нормальному закону. У разі збільшення обсягу вибірки середня вибіркового розподілу залишається незмінною, тобто не залежить від обсягу вибірки, проте стандартна похибка вибіркових середніх зменшується зі зростанням обсягу вибірки.

При достатньо великій кількості незалежних спостережень ймовірність того, що розбіжність між вибірковою та генеральною середніми не перевищить величини $t\mu$ визначається за формулою:

$$P(|\bar{x} - \tilde{x}| \leq t \cdot \mu) = 2\Phi(t),$$

де $\Phi(t)$ – значення функції Лапласа у точці *t*. Цю точку називають *коефіцієнтом довіри*.

Для формування вибірок потрібно визначити *основу генеральної сукупності*, тобто перелік її об’єктів. Об’єкти, що входять до складу генеральної сукупності, можна перенумерувати і відбирати об’єкти для вибірки за номерами.

Існують два основні типи вибірки. Після того, як об’єкт вийняли з генеральної сукупності для включення у вибірку, його після обстеження або повертають у генеральну сукупність, або не повертають. У відповідності з цим розрізняють вибірки з поверненням та вибірки без повернення.

4.2. Способи формування вибірок

Розрізняють наступні способи формування вибірок:
1) простий випадковий відбір;
2) механічний відбір;

3) типовий відбір;

4) серійний відбір.

Простий випадковий відбір передбачає формування вибірки жеребкуванням, без попереднього поділу генеральної сукупності на окремі групи. Цей спосіб вибіркового дослідження доцільно застосовувати, якщо між одиницями генеральної сукупності немає суттєвих відмінностей.

Жереб розігрують різноманітними способами, проте існує стандартний механізм жеребкування, що грунтується на використанні випадкових чисел, рівномірно розподілених на інтервалі (0;1). Сформувати вибіркову сукупність, до складу якої входять n елементів, із N елементів генеральної сукупності, можна за n послідовних кроків. Всі елементи генеральної сукупності повинні мати рівні шанси потрапити до вибіркової сукупності. На кожному кроці вибіркова сукупність поповнюється одним елементом з генеральної сукупності, який вибирається на підставі отриманого випадкового числа ξ.

Нехай A_i – подія, що полягає у виборі з генеральної сукупності i -ої за порядком слідування одиниці. Оскільки на кожному кроці до вибірки потрапляє лише один елемент генеральної сукупності, то події $A_1, A_2, ..., A_N$ утворюють повну групу, а ймовірність настання кожної події складає $1/N$. Інтервал (0;1) ділять на N рівних частин. Перша частина належить до події A_1, друга – A_2, і так далі. До вибіркової сукупності потрапляє одиниця спостереження генеральної сукупності з порядковим номером, що відповідає номеру частини, у якій розташоване значення випадкового числа ξ. У випадку формування простої вибірки за схемою неповерненої кулі із генеральної сукупності після кожного кроку виключається одиниця спостереження, що увійшла до вибірки.

Найпоширенішими методами отримання випадкових чисел, значення яких від 0 до 1 є рівноймовірними, є табличний та програмний методи. Перший полягає у використанні таблиць випадкових чисел, а другий – у реалізації відомих алгоритмів генерування випадкових чисел на комп’ютері.

При *механічному відборі* досліджують одиниці сукупності, розташовані на однаковій відстані та у певній послідовності серед впорядкованої генеральної сукупності. При цьому задається початок, крок відліку та обсяг вибірки. Початок відліку може бути встановлений довільно або жеребкуванням, а крок відліку $h = N/n$.

Типовий відбір передбачає формування вибіркової сукупності на основі попередньої структуризації генеральної сукупності і незалежного відбору елементів з кожної групи (типу). Типи можуть утворюватися штучно або використовуватися ті, що сформувалися природно. Розрізняють *типовий пропорційний* та *непропорційний відбори*.

При *типовому пропорційному відборі* обсяг вибірки і-ої типової групи визначається за формулою:

$$n_i = n \cdot \frac{N_i}{N}.$$

У цій формулі n_i – обсяг вибірки і-ої типової групи, n – загальний обсяг вибірки з генеральної сукупності, N_i – обсяг типової групи, N – обсяг генеральної сукупності.

Для непропорційної типової вибірки і-ої типової групи її чисельність визначається за формулою:

$$n_i = \frac{n}{K},$$

де К – кількість відокремлених груп.

Оптимальної вибірки досягають, вибираючи чисельність і-ої групи за формулою:

$$n_i = n \cdot \frac{N_i \sigma_i}{\sum N_i \sigma_i}.$$

Тут σ_i – середнє квадратичне відхилення для і-ої групи.

При *серійному відбор*і випадковим чином вибирають групи (серії) одиниць із генеральної сукупності. Серії можуть бути сформовані природно (регіони, партії продукції) або штучним способом.

Вибір певного способу формування вибіркової сукупності залежить від можливостей організації спостереження та його мети.

4.3 Вибіркові розподіли

Вибірковий метод дозволяє використовувати вибіркові статистики для оцінки різноманітних параметрів генеральної сукупності (генеральної середньої, стандартного відхилення, частки певних об’єктів у генеральній сукупності).

Будемо розглядати вибірки без повернення. Нехай вибірку обсягу n вийнято з генеральної сукупності, обсяг якої дорівнює N, велику кількість разів і кожного разу обчислюється певна вибіркова статистика. У результаті отримуємо ряд розподілу для цієї статистики, фіксуючи отримані її значення та частоти цих значень. Такий ряд розподілу для вибіркової характеристики називають *вибірковим розподілом*.

Приклад 4.1. Маємо скінченну генеральну сукупність, що складається з 6 чисел: 4, 8, 12, 16, 20,24. Побудуємо вибірковий розподіл для вибіркової середньої, використовуючи всі вибірки з 2 чисел.

Таблиця 4.1. Вибіркові середні.

Вибірка	Вибіркове середнє, $\overline{x} = \dfrac{x_1 + x_2}{2}$	Вибірка	Вибіркове середнє, $\overline{x} = \dfrac{x_1 + x_2}{2}$
4. 8	6	8, 24	16
4, 12	8	12, 16	14
4, 16	10	12, 20	16
4, 20	12	12, 24	18
4,24	14	16, 20	18
8, 12	10	16, 24	20
8, 16	12	20, 24	22
8, 20	14		

На основі цієї таблиці отримаємо вибірковий розподіл середніх.

Таблиця 4.2. Вибірковий розподіл середніх

Вибіркова середня, $\overline{x}$	6	8	10	12	14	16	18	20	22
Частота, f	1	1	2	2	3	2	2	1	1

У практиці статистичних досліджень найчастіше використовують вибіркові розподіли вибіркових середніх та вибіркових дисперсій або стандартних відхилень.

У подальшому будемо вважати, що генеральна сукупність має нормальний розподіл.

4.4. Вибірковий розподіл вибіркових середніх

Нехай нам потрібно з'ясувати середнє значення деякої ознаки для об'єктів генеральної сукупності (середню зарплату у галузі, середній вік населення країни тощо). Якщо генеральна сукупність розподілена за нормальним законом, то вибірковий розподіл вибіркових середніх також є нормальним. Навіть за відсутності нормального розподілу у генеральній сукупності для великих вибірок ($n \geq 30$) вибірковий розподіл випадкових середніх є близьким до нормального. Це дає змогу використовувати формули, що виконуються для нормального розподілу.

За вибірковим розподілом ми можемо обчислити середнє значення вибіркових середніх:

$$m(\overline{x}) = \frac{\sum \overline{x} \cdot f}{\sum f}. \tag{4.1}$$

У цій рівності $m(\overline{x})$ – середнє значення вибіркових середніх.

Якщо генеральна сукупність розподілена за нормальним законом, то $m(\overline{x})$ наближено дорівнює середньому значенню для генеральної сукупності. Це твердження справедливе для випадкових вибірок. Тоді середня, отримана за вибірковим розподілом, є незміщеною оцінкою генеральної середньої.

Приклад 4.2. Знайдемо $m(\overline{x})$ для прикладу 4.1. Маємо:

$$m(\overline{x}) = \frac{6 \cdot 1 + 8 \cdot 1 + 10 \cdot 2 + 12 \cdot 2 + 14 \cdot 3 + 16 \cdot 2 + 18 \cdot 2 + 20 \cdot 1 + 22 \cdot 1}{1 + 1 + 2 + 2 + 3 + 2 + 2 + 1 + 1} = 14.$$

У цьому прикладі середня вибіркових середніх $m(\overline{x})$ співпадає з середньою μ для генеральної сукупності:

$$\mu = \frac{4 + 8 + 12 + 16 + 20 + 24}{6} = 14.$$

На практиці для знаходження статистичної оцінки генеральної середньої не будують велику кількість вибірок з подальшим знаходженням їх вибіркових середніх та обчисленням їх середньої величини у якості оцінки генеральної середньої. Найчастіше маємо дані однієї вибірки, для якої можемо обчислити вибіркове середнє і використати його як оцінку генеральної середньої: $\tilde{\mu} = \overline{x}$, де $\tilde{\mu}$ – оцінка генеральної середньої. Похибку такої оцінки можна виразити через дисперсію вибіркового розподілу середніх. Стандартне відхилення *SE* вибіркового розподілу середніх (його стандартну похибку) обчислюють за формулою:

$$SE_{\overline{x}} = \sqrt{\frac{\sum f\left(\overline{x} - m(\overline{x})\right)^2}{\sum f}} = \sqrt{\left(\frac{\sum f \cdot \overline{x}^2}{\sum f} - \left(m(\overline{x})\right)^2\right)}. \tag{4.2}$$

Якщо генеральна сукупність має нормальний розподіл, то стандартну похибку вибіркового розподілу вибіркових середніх можна визначити за формулою:

$$SE_{\bar{x}} = \sqrt{\frac{(N-n)\sigma^2}{(N-1)n}}. \qquad (4.3)$$

Тут σ^2 – дисперсія генеральної сукупності.

Якщо генеральна сукупність є великою у порівнянні з обсягом вибірки (якщо $\frac{n}{N} \le 0,05$), то $\sqrt{\frac{(N-n)}{(N-1)}} \approx 1$ і стандартна похибка набуває вигляду:

$$SE_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}}. \qquad (4.4)$$

Якщо ми будемо змінювати обсяг вибірки, то середня вибіркового розподілу середніх не зміниться, а стандартна похибка $SE_{\bar{x}}$ зменшується при збільшенні обсягу вибірки.

При обчисленні стандартної похибки вибіркових середніх ми вважали, що нам відома генеральна дисперсія σ^2, хоча фактично нам ця величина невідома. Тому потрібно розв'язати задачу оцінки генеральної дисперсії за вибіркою.

4.5. Вибірковий розподіл вибіркової дисперсії

Вибірковий розподіл вибіркової дисперсії можна вивчати з допомогою методу, що був застосований для аналізу вибіркового розподілу вибіркових середніх, обчислюючи дисперсію кожної вибірки. Проте слід враховувати, що вибірковий розподіл вибіркових дисперсій не є нормальним. Якщо генеральна сукупність має нормальний розподіл, то вибірковий розподіл вибіркової дисперсії має розподіл χ^2.

Розглянемо випадок скінченної генеральної сукупності. Нехай s^2 – вибіркова дисперсія. Її середнє значення

$$m(s^2) = \frac{N}{N-1} \cdot \frac{(n-1)\sigma^2}{n}. \qquad (4.5)$$

Звідси можна визначити генеральну дисперсію σ^2.

При знаходженні незміщеної оцінки генеральної дисперсії, як і у випадку генерального середнього, можемо мати лише одну вибірку і таку оцінку доведеться знаходити лише на основі вибіркової дисперсії цієї вибірки. Позначимо оцінку генеральної дисперсії $\tilde{\sigma}^2$. Тоді отримуємо:

$$\tilde{\sigma}^2 = \frac{N-1}{N} \cdot \frac{ns^2}{n-1}, \; s^2 = \frac{\sum\left(x-\bar{x}\right)^2}{n}. \tag{4.6}$$

Приклад 4.3. За умови прикладу 4.1 побудувати вибірковий розподіл на основі вибірок з 2 елементів та розрахувати вибіркову дисперсію як статистику для кожної вибірки.

Таблиця 4.4. Вибіркові дисперсії для вибірок з 2 елементів

Вибірка	Вибіркова дисперсія, $s^2 = \frac{\left(x_1-\bar{x}\right)^2+\left(x_2-\bar{x}\right)^2}{2}$	Вибірка	Вибіркова дисперсія, $s^2 = \frac{\left(x_1-\bar{x}\right)^2+\left(x_2-\bar{x}\right)^2}{2}$
4. 8	4	8, 24	64
4, 12	16	12, 16	4
4, 16	36	12, 20	16
4, 20	64	12, 24	36
4,24	100	16, 20	4
8, 12	4	16, 24	16
8, 16	16	20, 24	4
8, 20	36		

Побудуємо вибірковий розподіл вибіркових дисперсій.

Таблиця 4.5. Вибірковий розподіл вибіркових дисперсій

Вибіркова дисперсія, s^2	4	16	36	64	100
Частота, f	5	4	3	2	1

Знайдемо середнє значення отриманого ряду розподілу для вибіркових дисперсій:

$$m\left(s^2\right) = \frac{\sum f \cdot s^2}{\sum f} = \frac{5\cdot4+4\cdot16+3\cdot36+2\cdot64+1\cdot100}{5+4+3+2+1} = 28.$$

Використовуючи формулу (4.5), знайдемо генеральну дисперсію σ^2 :

$$\sigma^2 = \frac{N-1}{N} \cdot \frac{n}{n-1} \cdot m\left(s^2\right) = \frac{6-1}{6} \cdot \frac{2}{2-1} \cdot 28 = 46,6 .$$

Оскільки нам відомі всі елементи генеральної сукупності, то ми можемо безпосередньо обчислити генеральну дисперсію (генеральна середня $\mu = 14$):

$$\sigma^2 = \frac{\left(x-\mu\right)^2}{n} = \frac{1}{6} \cdot ((4-14)^2 + \left(8-14\right)^2 + \left(12-14\right)^2 + \left(16-14\right)^2 +$$
$$+\left(20-14\right)^2 + \left(24-14\right)^2) = 46,6.$$

Отримана генеральна дисперсія дорівнює значенню, отриманому з використанням вибіркового розподілу вибіркових дисперсій.

Розглянемо випадок великої генеральної сукупності. Для такої сукупності $\frac{N}{N-1} \approx 1$, тоді вираз (4.6) для незміщеної оцінки генеральної дисперсії набуває вигляду:

$$\tilde{\sigma}^2 = \frac{ns^2}{n-1}. \tag{4.7}$$

Якщо генеральна сукупність має нормальний закон розподілу, то вибірковий розподіл величини $\frac{ns^2}{\sigma^2}$ є розподілом χ^2 з $n-1$ ступенями вільності, значення функції цього розподілу наведені у статистичних таблицях.

Формулу (4.7) можна використати для отримання оцінки стандартної похибки вибіркового розподілу вибіркових середніх. Раніше ця похибка визначалася за формулою (4.4):

$$SE_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} .$$

Щоб використати цю формулу, потрібно знати генеральну дисперсію σ^2 . Якщо нам її величина невідома, то можна використати її оцінку (4.7). Підставивши її замість σ^2 у (4.4), отримуємо оцінку стандартної похибки вибіркового розподілу вибіркових середніх у вигляді:

$$\widetilde{SE_{\overline{x}}} = \sqrt{\frac{ns^2}{n(n-1)}} = \sqrt{\frac{s^2}{n-1}}\,, \tag{4.8}$$

де $s^2 = \frac{\sum(x-\overline{x})^2}{n}$.