УДК 576.8:57.081.17:311.2 DOI https://doi.org/10.26661/2410-0943-2019-1-07

Визначення раціонального обсягу вибірки в паразитологічних дослідженнях бутстреп-методом

[1]Швидка С. П., [1]Левчук С. А., [2]Сарабєєва Є. В.

ORCID: 0000-0002-9773-278X

[1]*Запорізький національний університет, Україна*

[2]*Запорізька гімназія № 28, Україна*

svetlana.shvydka@gmail.com; levchukser65@gmail.com

Ключові слова:
бутстреп, довірчий інтервал, точність, агрегований розподіл, паразити риб

Висока трудомісткість і вартість паразитологічних досліджень вимагає використання раціонального обсягу вибірки як балансу між інформативністю даних із малих вибірок та непотрібними витратами на отримання надлишкових даних. Агрегований характер розподілу паразитів у популяціях хазяїв створює принципові труднощі при розрахунку інтервальних оцінок вибіркових характеристик методами класичної статистики. У таких ситуаціях доцільно застосовувати непараметричні методи і ресамплінг. Запропоновані практичні рекомендації дозволяють визначати раціональний обсяг вибірки для обчислення інтервальних оцінок статистичних характеристик із заданою точністю з використанням методу Bag of Little Bootstraps. Точність визначається як довірчий інтервал, такий, що оцінка середнього значення повинна бути в межах певної величини істинного середнього із заданою ймовірністю. Підхід проілюстровано на прикладі визначення раціонального обсягу вибірки для обчислення інтервальних оцінок вибіркового середнього значення чисельності паразитів роду *Ligophorus*, *L. llewellyni* та *L. pilengas.* Для кожного з видів паразитів обсяг емпіричних вибірок дорівнював 224 елементам. Результати дослідження показали, що обсяг вибірки значною мірою залежить від рівня точності та параметра агрегації. Виявлено, що для вибірок обсягом 45 елементів і більше довжина довірчого інтервалу дорівнює емпіричному середньому або є меншою; вибірки з чисельністю у діапазоні 20 – 35 екземплярів недооцінюють значення вибіркового середнього, а малі вибірки (n=10) призводять до ненадійних оцінок. Схожі результати отримані в попередніх дослідженнях Маркес і Кабрал. Запропонований підхід стане в нагоді паразитологам при плануванні дизайну вибірки та допоможе визначити рівень точності вибіркового середнього значення у дослідженнях.

Determination of rational sample size in parasitological studies by bootstrap method

[1]Shvydka S. P., [1]Levchuk S. A., [2]Sarabeeva Ye. V.

[1]*Zaporizhzhia National University, Ukraine*

[2]*Zaporizhzhia high school №28, Ukraine*

Key words:
bootstrap, confidence interval, precision, aggregated distribution, fish parasites

Sampling of field material in parasitological studies is a highly time consuming procedure, which successful implementation also requires substantial financial and human resources. Therefore it is important to determine the rational sample size as a balance between the informative of the data from the small samples and the excess costs for collecting the unnecessary data. Due to aggregated distribution of parasites in populations of hosts it is difficult to calculate the confidence intervals of the statistical characteristics using the methods of classical statistics. In such situations it is need to use the non-parametric methods and resampling. The study offers practical recommendations to determine the rational sample size for calculation the interval estimates of the statistical parameters and the precision using the Bag of Little Bootstraps. The precision is defined as a confidence interval such that the estimate of the mean should be within some value of the true mean. The approach is illustrated on the example of monogenean parasites of *Ligophorus*

llewellyni and *L. pilengas* from the so-iuy mullet. The initial data set included 224 elements for each parasite species. The results showed that the level of precision and parameter of aggregation are strongly affected by the sample size. It was found that the width of the confidence interval was equal or less of the empirical mean for samples more than 45 elements. The mean abundance is systematically underestimated for samples with 20–35 individuals. The small samples (n=10) have led to the unreliable estimates. The similar results were obtained in studies of Marques and Cabral. The proposed here approach will relief the sampling plan design and will help researchers to define the rational sample size and the precision level for the estimated mean abundance in parasitological studies.

Вступ

Обсяг вибірки є одним із визначальних елементів планування статистичного дослідження в будь-якій галузі науки. Занадто велика вибірка призводить до невиправданих витрат та неефективного використання ресурсів, а недостатня – до незадовільної якості результатів дослідження. З огляду на це, питання розробки підходів до визначення раціонального обсягу вибірки, яка б гарантувала якість результатів спостереження при плануванні вибіркового дослідження, залишається актуальним і сьогодні.

Існує декілька стратегій для визначення обсягу вибірки[1]. Першим підходом є використання усієї сукупності даних, що виключає помилку вибірки та надає дані про всі особи у популяції. Цей метод є прийнятним лише для невеликих популяцій (наприклад, 200 або менше). Другий підхід полягає у використанні такого самого обсягу вибірки, як і в попередніх дослідженнях, схожих на ті, що плануються. У такому разі огляд літератури може надати рекомендації щодо «типової» вибірки, але при цьому необхідно проаналізувати коректність процедури визначення розміру вибірки в попередніх дослідженнях. Третій спосіб передбачає використання опублікованих таблиць[1,2], які містять інформацію про обсяг вибірки для заданої множини критеріїв (розміра популяції, точність, довірчий інтервал, закон розподілу). Використання четвертого підходу, а саме розрахунку обсягу вибірки за допомогою математичних формул[3–5], вимагає виконання передумови про певний теоретичний закон розподілу результатів спостереження. Нарешті, п'ятим шляхом є застосування некласичних методів статистики, що суттєво спираються на можливості сучасної комп'ютерної техніки. Такі методи базуються на ідеї генерації штучних вибірок на основі результатів спостереження. Подальший аналіз генерованих даних дозволяє отримати задовільну статистичну інформацію.

Проблема аналізу паразитологічних даних ускладняється тим, що у випадку природних інфекцій паразити зазвичай демонструють агрегований розподіл, тобто більшість хазяїв містять невелику кількість паразитів і окремі особи – занадто багато паразитів[6–8]. У більшості випадків такий розподіл може бути математично змодельованим за допомогою негативного біноміального розподілу $NB(m,k)$, де m – вибіркове середнє, k – параметр, що характеризує агрегацію[7]. Агрегований характер розподілу паразитів впливає на вибіркове середнє та його інтервальну оцінку[9], тому при визначенні раціонального обсягу вибірки слід знайти компроміс між репрезентативністю паразитологічних даних із малих вибірок та непотрібними витратами на отримання надлишкових даних. Попередні дослідження[5,10] продемонстрували, що найбільш потужними інструментами визначення оптимального обсягу вибірки для агрегованих даних є Монте-Карло симуляції та бутстреп-процедури.

Метою роботи є надання практичних рекомендацій щодо оптимізації визначення обсягу вибірки в паразитологічних дослідженнях із використанням бутстреп-моделювання.

Традиційний метод бутстрепа[11] вимагає, щоб вибірки бутстрепа були такого самого розміру, що й оригінальна вибірка. На відміну від нього, метод Bag of Little Bootstraps (BLB)[12] формує вибірки бутстрепа з декількох підвибірок або підмножин

вибірки, яка отримана внаслідок експерименту або спостереження. Згідно з BLB-методом, спочатку з оригінальної вибірки будують кілька підвибірок однакового обсягу, для яких потім застосовують традиційний бутстреп-метод, отримуючи кілька бутстреп-характеристик. Отже, на відміну від традиційного бутстрепа, який дає одне-єдине значення статистичної характеристики, BLB-метод дозволяє отримати бутстреп-характеристики для кожної з підвибірок. Зауважимо, що підвибірки з оригінальної вибірки містять менше унікальних спостережень, а отже, менше інформації, ніж вибірки традиційного бутстрепа, побудовані на основі оригінального набору даних. У BLB-методі вказана проблема вирішується шляхом осереднення усіх бутстреп-характеристик для підвибірок однакового обсягу.

Для практичної реалізації бутстреп-моделювання досить зручним є використання статистичного середовища R (https://www.r-project.org), яке стрімко набуває популярності при проведенні статистичних обчислень.

Матеріали та методи

Згідно з BLB-методом, визначення раціонального обсягу вибірки для знаходження обґрунтованої інтервальної оцінки вибіркової характеристики може бути виконане за схемою:

1) з емпірчної вибірки даних випадковим чином утворити підвибірку обсягом n. Значення n має бути меншим ніж обсяг початкової емпіричної вибірки або дорівнюватися цьому обсягу;

2) для створеної підвибірки побудувати B вибірок бутстрепа та обчислити бутстреп-характеристику вибіркового параметра;

3) кроки 1 і 2 повторити N раз;

4) для отриманих N бутстреп-характеристик значення вибіркового параметра виконати ранжування за збільшенням;

5) для побудови 95% довірчого інтервалу знайти значення вибіркового параметра, які являють собою 2,5 та 97,5 процентилі. Це будуть ліва та права межі 95% довірчого інтервалу;

6) повторити кроки 1 – 5 для різних значень n, зважаючи на цілі дослідження;

7) проаналізувати зміну довжини довірчого інтервалу зі зміною величини n і обрати раціональний обсяг вибірки як баланс між задовільною для практичного застосування інтервальною оцінкою вибіркового параметра та допустимим рівнем невизначеності такої оцінки;

8) використовуючи межі довірчого інтервалу, обчислити точність.

Алгоритм процесу отримання довірчого інтервалу для бутстреп-характеристики наведено на рисунку 1.

Запропонований підхід проілюстровано на прикладі визначення раціонального обсягу вибірки для обчислення обґрунтованих інтервальних оцінок вибіркового середнього значення ектопаразитичних моногенетичних сисунів роду *Ligophorus*, *L. llewellyni* та *L. pilengas*, від кефалі піленгаса, зібраної з Японського моря в період 2004 – 2005 років у межах виконання проєкту ІНТАС № 03-51-5998. Для кожного з видів паразитів обсяг емпіричних вибірок складав 224 елементи. На основі отриманих вибірок були виконані розрахунки для n від 10 до 100 з кроком 5 та таких параметрів: B=1000, N=10000. Розрахунки проведені в середовищі R (version 3.6.1, R Development Core Team, 2019) з використанням такого коду:

```
R> install.packages("MASS") # інсталяція пакета
R>library(MASS) # завантаження пакета
R> outp = matrix(nrow=N, ncol=2) # створення матриці для запису бутстреп вибіркових середніх значень
R> for(i in 1:N){ # цикл, що дозволяє обчислювати вибіркові середні значення бутстрепа для N різних підвибірок
R> subsample<-c(sample(data,10)) # створення підвибірки обсягом 10 елементів з емпіричних даних
R> boot <-numeric(B) # створення вектору для зберігання вибіркових середніх значень для кожної вибірки бутстрепа
R> for (j in 1:B) boot[j] <- (mean(sample(subsample,replace=T))) # бутстреп
R> mean(boot) # бутсреп вибіркове середнє значення для підвибірки
R> outp[i,] = c(i,mean(boot)) # запис N бутстреп вибіркових середніх значень
R> }
```

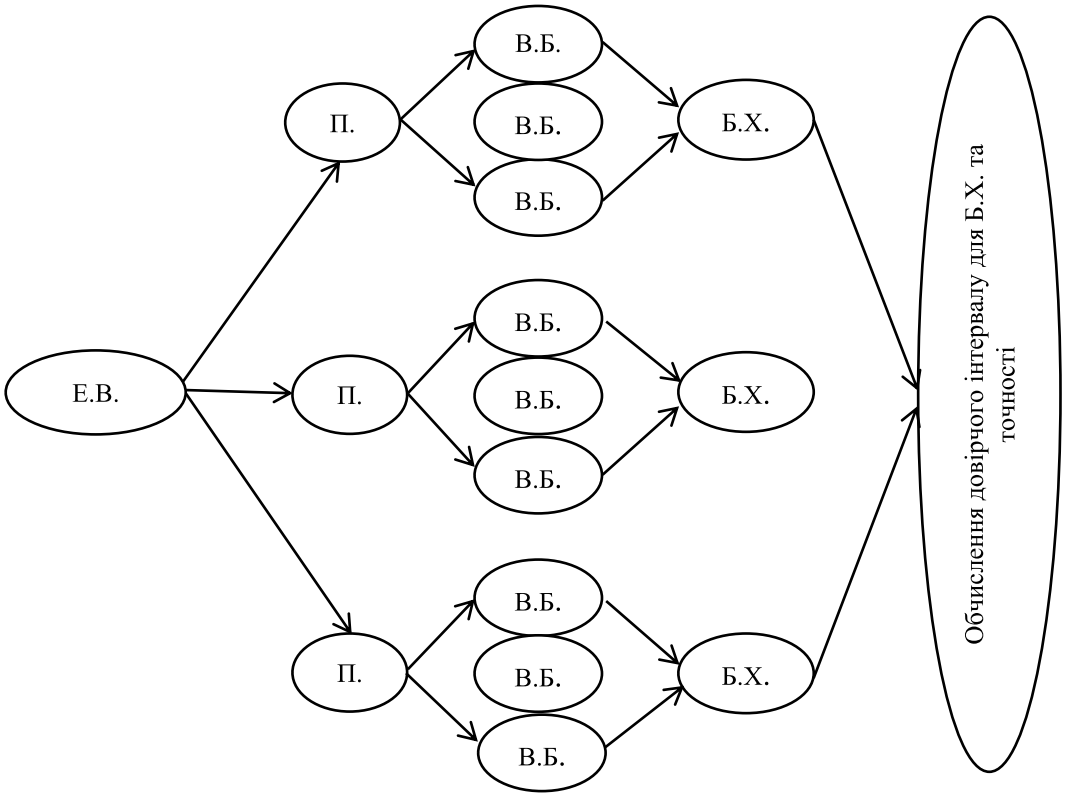

Рис. 1. Алгоритм процесу отримання довірчого інтервалу для бутстреп-характеристики (Б.Х.); Е.В. – емпірична вибірка, П. – підвибірка, В.Б. – вибірка бутстрепа

Для порівняння результатів, отриманих бутстреп-моделюванням, були використані методи класичної статистики, а саме розрахунок обсягу вибірки за допомогою математичної формули[3]:

$$n = \left(\frac{Z_{\alpha/2}}{D}\right)^2 \left(\frac{1}{m} + \frac{1}{k}\right), \qquad (1)$$

де $Z_{\alpha/2}$ – квартиль нормального розподілу, D – точність, m – вибіркове середнє значення, k – параметр, що характеризує агрегацію. За M. G. Karandinos[3], одним із найпоширеніших шляхів до визначення точності D є визначення її як довірчого інтервалу, такого, що оцінка середнього значення повинна бути в межах певної величини істинного середнього значення:

$$CI/2 = D \times m. \qquad (2)$$

Результати

Розподіл чисельності обох досліджених видів моногеней у популяції хазяїна мав виражений агрегований характер (відношення дисперсії до середнього значення значно більше за 1 (табл. 1)). Перевірка закону розподілу з використанням тесту χ^2 показала, що обидві вибірки паразитів підпорядковуються негативному біноміальному розподілу (табл. 1).

Таблиця 1 – Статистичні характеристики емпіричних даних

Вид моногеней	Вибіркове середнє	Дисперсія	k (p-значення)
Ligophorus llewellyni	42.83	5921.869	0.45 (0.28)
L. pilengas	22.19	1852.44	0.34 (0.11)

Результати бутстреп-моделювання вибіркового середнього значення надані на рисунку 2 у вигляді діаграм розмаху для вибірок обсягом від 10 до 50 елементів. Із діаграм видно, що для малих вибірок розподіл бутстреп-характеристик вибіркового середнього має правосторонню асиметрію. Для таких несиметричних розподілів медіана дає більш точну характеристику ознаки, і тому її

використовують замість вибіркового середнього значення. Для обох видів паразитів вибірки обсягом менше ніж 35 елементів недооцінюють значення вибіркового середнього.

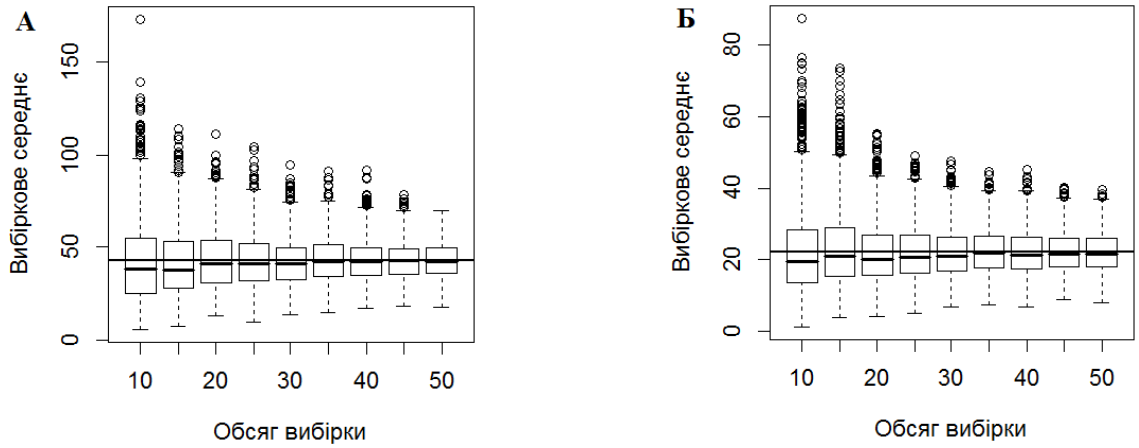

Рис. 2. Розподіл вибіркового середнього значення, отриманий для вибірок різного обсягу за допомогою BLB-методу; *L. llewellyni* (А) та *L. pilengas* (Б). Прямокутник включає значення між першим і третім квартилями. Вертикальні лінії (вуса), докреслені до прямокутника, відображають мінливість значень за межами верхнього й нижнього квартилів, мінімальне та максимальне значення, позначені рисками, викиди – кругами. Лінією в прямокутнику позначена медіана. Пряма лінія – емпіричне значення вибіркового середнього

На рисунку 3 наведено 95% довірчі інтервали для вибіркового середнього значення. Як видно, довірчі інтервали є несиметричними, що відповідає асиметрії вибіркового середнього чисельності паразитів. При збільшенні розміру вибірки інтервали помітно звужуються. Найшвидше зменшення довжини довірчого інтервалу спостерігалося для невеликих вибірок (< 40 елементів), тоді як подальше збільшення обсягу вибірок призвело до сповільненого зменшення довжини довірчого інтервалу. За умови збільшення обсягу вибірки на однакову величину для малих та великих вибірок, довжина довірчого інтервалу змінюється по-різному. Наприклад, збільшення обсягу вибірки з 20 до 40 та з 50 до 70 елементів зменшує довжину довірчого інтервалу на величину 0,5 × вибіркове середнє та 0,17 × вибіркове середнє, відповідно. Збільшення обсягу вибірки з 70 до 100 елементів не призводить до помітного звуження довжини довірчого інтервалу, зменшуючи його лише на величину 0,18 × вибіркове середнє.

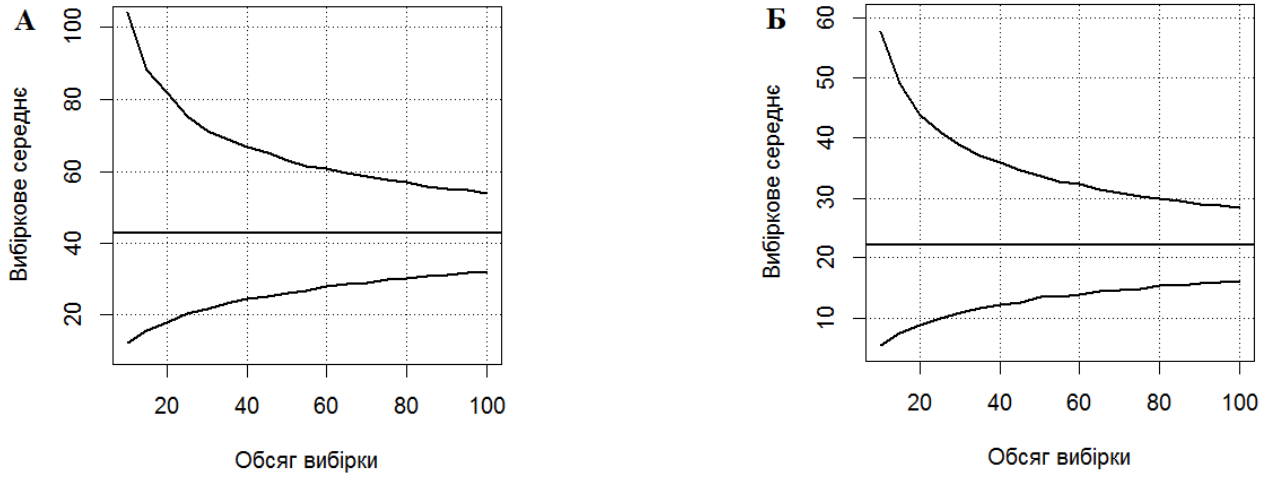

Рис. 3. Вибіркові середні значення та їхні 95% довірчі інтервали (пряма лінія – емпіричне значення вибіркового середнього); *L. llewellyni* (А) та *L. pilengas* (Б)

Криві зміни обсягу вибірки від точності D, побудовані на рис. 4, ілюструють швидке збільшення точності для малих вибірок, що містять від 10 до 40 елементів, та значно сповільнений її ріст для великих вибірок (> 40 елементів). За умови фіксованої точності саме параметр агрегації k має визначальний вплив на обсяг вибірки. Зі зменшенням величини k, інакше кажучи, при збільшенні рівня агрегації, слід розглядати вибірки більшого обсягу. Для однакових значень точності, обчислення, виконані за формулою (1), добре узгоджуються з результатами бутстреп-моделювання (рис. 4).

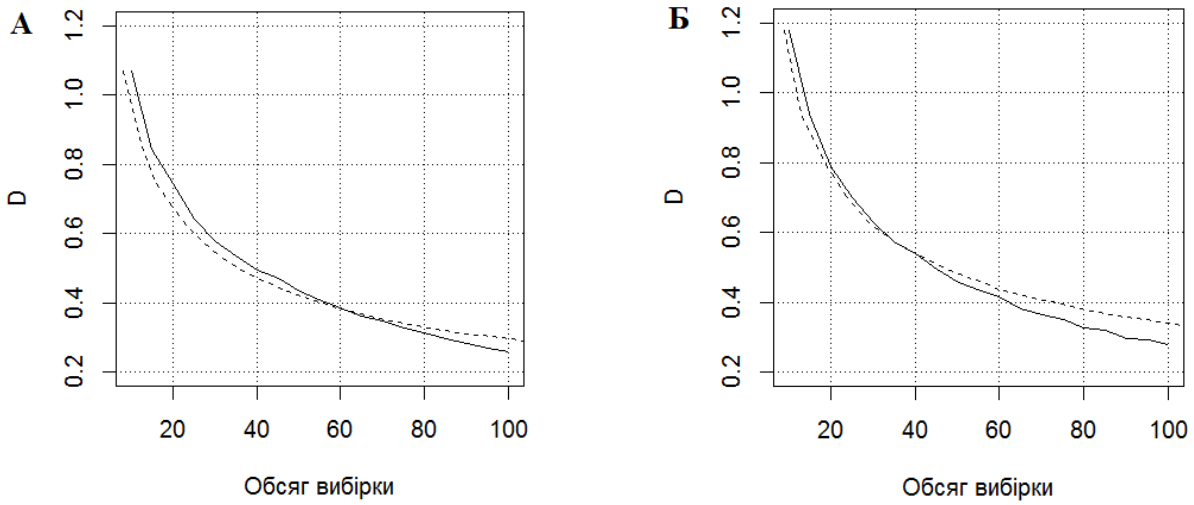

Рис. 4. Вплив точності D на обсяг вибірки для оцінки вибіркового середнього значення для моногеней *L. llewellyni* (А) та *L. pilengas* (Б). Неперервна лінія розрахована на основі емпіричних даних з використанням бутстреп-моделювання та формули (2), переривчаста лінія – за формулою (1)

Обговорення

Проведене комп'ютерне моделювання виявило ефективність використання бутстреп-процедури для задачі визначення раціонального обсягу вибірки для обчислення обґрунтованих інтервальних оцінок вибіркових характеристик з прийнятним рівнем точності в паразитологічних дослідженнях. Обсяг вибірки залежить від двох факторів: необхідного рівня точності та параметра агрегації. За інших однакових умов, для даних з високим рівнем агрегації обсяг вибірки має бути більшим для досягнення високого рівня точності. Це пояснюється рідкістю спостереження сильно заражених хазяїв і тому, швидше за все, ймовірність виявлення їх у вибірках невеликого обсягу є незначною[8].

Інтервальні оцінки статистичних параметрів дозволяють дослідникам оцінити значимість отриманих результатів, а довжина довірчих інтервалів – надати інформацію про точність вибіркових характеристик. У ситуаціях з агрегованими наборами даних доцільно застосовувати непараметричні методи і ресамплінг, що намагаються зрозуміти розподіл емпіричних даних безпосередньо у ході обчислень і оцінити параметри невідомих та складних законів розподілу. Бутстреп-моделювання дозволяє проводити детальний аналіз емпіричного матеріалу таким чином, що кожний окремий елемент з експериментальної вибірки дає свій вагомий внесок у формування кінцевого результату. На відміну від методів класичної статистики, моделювання дозволяє не втратити статистичних особливостей емпіричних даних, тоді як перші їх згладжують. Особливо це є актуальним для даних з асиметричним розподілом, якими є агреговані дані.

Швидке збільшення точності з ростом обсягу вибірки вказує на зниження рівня невизначеності, а отже, означає, що вибірка стає репрезентативною для отримання допустимих інтервальних оцінок. Подальше повільне лінійне збільшення точності зі збільшенням обсягу вибірки пояснюється високою загальною кількістю ненульових значень у вибірках. Невеликі вибірки є

достатніми у випадках, коли менший рівень точності є прийнятним, а також для даних із невеликою агрегацією[4, 5].

Для обох видів паразитів результати, отримані в роботі, добре узгоджуються з попередніми дослідженнями. Із використанням методу Монте-Карло J. F. Marques і H. N. Cabral[10] з'ясували, що невеликі вибірки (< 40 елементів) систематично занижують значення показників середньої чисельності та середньої інтенсивності паразитів. Цей факт пояснюється наявністю правосторонньої асиметрії у розподілі паразитологічних даних із великою кількістю нулів, що означають відсутність паразита, та незначною кількістю великих даних, які характеризують сильно заражених риб. У роботі[5] показано, що для вибірок обсягом 80 елементів і більше величина довірчого інтервалу дорівнює емпіричному середньому значенню або є меншою; вибірки з чисельністю в діапазоні 25-40 екземплярів недооцінюють значення середнього, а малі вибірки (n=10) призводять до ненадійних оцінок.

Застосований підхід, по-перше, дає змогу уникнути необхідності робити не завжди обґрунтовані припущення про певний закон розподілу випадкової величини, що досліджується. Процес розрахунку не привносить в отримані статистики жодної зайвої інформації та дозволяє досить ретельно аналізувати статистичні дані, що представлені складними законами розподілу, не допускаючи втрати інформації. По-друге, алгоритм процесу дозволяє прослідкувати зміну розподілу вибіркових характеристик при використанні практично необмеженої кількості штучних повторних вибірок, отриманих в однакових умовах. Зрештою, реалізація підходу не є складною і звільняє від необхідності пошуку математичних формул та критеріїв, які є найкращими для статистичної обробки конкретних даних.

Висновки

Запропонований алгоритм побудови інтервальних оцінок вибіркових параметрів стане в нагоді паразитологам при плануванні перспективного дизайну вибірки та допоможе дослідникам зрозуміти рівень точності вибіркових характеристик у паразитологічних дослідженнях.

Розглянутий підхід до визначення раціонального обсягу вибірки дозволяє проводити обробку експериментальних даних, що представлені вибірками малого обсягу та складними законами розподілу, а також у випадках, коли методи класичної статистики не можуть бути застосовані. При аналізі невеликих вибірок, метод бутстрепа буде корисним для обчислення описової статистики з відповідними довірчими інтервалами. Такий підхід є доцільним у ситуаціях, коли малі вибірки є неминучими, наприклад, при роботі з видами, які перебувають на межі вимирання. Сучасні інформаційно-обчислювальні технології допоможуть реалізувати практичні рекомендації і дадуть наочні результати, які легко піддаються інтерпретації.

Подяки

Автори висловлюють подяку доценту кафедри біології лісу, мисливствознавства та іхтіології Запорізького національного університету Сарабєєву В. Л. за допомогу та слушні поради щодо постановки наукової проблеми, а також за люб'язно надані емпіричні дані.

Робота Швидкої С. П. частково підтримана Національною стипендіальною програмою Словацької Республіки (SAIA – National Scholarship Programme of the Slovak Republic), проєкт № ID 24637.

Література

(1) Israel, G. D. *Determining Sample Size*; PEOD6; 1992.

(2) Курочкин, Ю. В. *Методическое пособие по паразитологическому инспектированию морских рыб*; ТИНРО: Владивосток, 1979.

(3) Karandinos, M. G. Optimum Sample Size and Comments on Some Published Formulae. *Bull. Entomol. Soc. Am.* **1976**, *22* (4), 417–421 DOI: 10.1093/besa/22.4.417.

(4) Opit, G. P.; Throne, J. E.; Flinn, P. W. Sampling Plans for the Psocids Liposcelis entomophila

and Liposcelis decolor (Psocoptera: Liposcelididae) in Steel Bins Containing Wheat. *J. Econ. Entomol.* **2009**, *102* (4), 1714–1722 DOI: 10.1603/029.102.0440.

(5) Shvydka, S.; Sarabeev, V.; Estruch, V. D.; Cadarso-Suárez, C. Optimum sample size to estimate mean parasite abundance in fish parasite surveys. *Helminthologia* **2018**, *55* (1), 52–59 DOI: 10.1515/helm-2017-0054.

(6) Anderson, R. M.; Gordon, D. M. Processes influencing the distribution of parasite numbers within host populations with special emphasis on parasite-induced host mortalities. *Parasitology* **1982**, *85 (Pt 2)*, 373–398.

(7) Shaw, D. J.; Dobson, A. P. Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. *Parasitology* **1995**, *111 Suppl*, S111-27.

(8) Poulin, R. Explaining variability in parasite aggregation levels among host samples. *Parasitology* **2013**, *140*, 541–546 DOI: 10.1017/S0031182012002053.

(9) Rózsa, L.; Reiczigel, J.; Majoros, G. Quantifying parasites in samples of hosts. *J. Parasitol.* **2000**, *86* (2), 228–232 DOI: 10.1645/0022-3395(2000)086[0228:QPISOH]2.0.CO;2.

(10) Marques, J. F.; Cabral, H. N. Effects of sample size on fish parasite prevalence, mean abundance and mean intensity estimates. *J. Appl. Ichthyol.* **2007**, *23* (2), 158–162 DOI: 10.1111/j.1439-0426.2006.00823.x.

(11) Efron, B.; Tibshirani, R. J. *An introduction to the bootstrap. Monographs on Statistics and Applied Probability, No. 57. Chapman and Hall, London, 436 p*; Chapman and Hall: London, 1993; Vol. 57.

(12) Kleiner, A.; Talwalkar, A.; Sarkar, P.; Jordan, M. I. A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **2014**, *76* (4), 795–816 DOI: 10.1111/rssb.12050.