МЕТОДИ МАТЕМАТИЧНОГО МОДЕЛЮВАННЯ

В.О. СТРОЄВА, к.ф.-м.н., доцент
М. НАКОПІЯ, магістр
Дніпровський державний технічний університет, м. Кам'янське

Застосування непараметричних методів до статистичного оцінювання вибіркових характеристик

Приведено результати застосування непараметричних методів до статистичного оцінювання вибіркових характеристик. Розглянуті основні ідеї методів чисельного ресамплінгу та показані їх переваги при обробці екологічних спостережень

Вступ

Математичне моделювання одержало широке поширення в різних галузях науки: механіці, фізиці, статистиці, медицині, біології, у тому числі й в екології. Багато таких задач пов'язано з обробкою даних методами статистичних досліджень, які, в свою чергу, неминуче засновані на обчисленнях. Ефективність і результативність їх реалізації повинні бути найбільш важливими і об'єктивними аргументами у прийнятті рішень. Застосування комп'ютерної техніки суттєво змінило концепцію обробки даних. Як зазначалося в [1], саме комп'ютерні технології сприяли розвитку нового класу альтернативних комп'ютерно-інтенсивних (computer-intensive) технологій, що включають рандомізацію, бутстрап і методи Монте-Карло, об'єднані загальним терміном "чисельний ресамплінг — resampling".

Відомо, що обробка спостережень параметричними методами ґрунтується на цілому ряді апріорних припущень, таких як незалежність вимірювань та їх помилок, однорідність дисперсій, нормальність розподілу і інше, якщо вони вірні, то тести мають безсумнівну надійність. У той же час можливі відхилення від цих припущень, характерні, наприклад, для екологічних даних, можуть серйозно вплинути на обґрунтованість кінцевих висновків. Отже, сучасною альтернативою параметричних методів є моделювання емпіричного розподілу даних з використанням методів генерації повторних вибірок (чисельного ресамплінгу) [2]. Ці технології не вимагають ніякої апріорної інформації про закон розподілу досліджуваної випадкової величини. Замість цього вони виконують багаторазову обробку різних емпіричних даних.

Параметричні методи оцінки надійних інтервалів у багатьох випадках характеризуються надійністю і теоретичною обґрунтованістю. Однак на практиці ці методи доводиться застосовувати і в тих випадках, коли спостереження не цілком відповідають основним постулатам про властивості випадкової величини, що надає обчисленням свідомо наближений і навіть некоректний характер. Такі можливі відхилення, характерні для екологічних або економічних даних, можуть серйозно вплинути на обґрунтованість кінцевих висновків: привести до зміщення оцінок, надійних границь і коефіцієнтів зв'язку. Часто ці порушення буває важко виявити по обмеженому числу спостережень і вони небезпечні саме цією непомітністю. При явній відміні розподілу від Гауссіана коректне довірче оцінювання і перевірка гіпотез про параметри переростає в складну проблему. У таких випадках розумно взагалі відмовитися від стандартної нормальної моделі і застосовувати непараметричні методи, засновані на ідеях Монте-Карло [3, 4].

Непараметричними називають такі методи статистики, що не залежать від якогось розподілу з теоретичного сімейства та не використовують його властивості. Вони спираються лише на припущення, що випадкова величина X незалежна і тотожно розподілена. У зв'язку зі своєю простотою та універсальністю непараметричні моделі набули широкого застосування і склали ефективну конкуренцію традиційним параметричним методам.

Спочатку непараметричні методи призначалися для перевірки статистичних гіпотез. У тому випадку, коли немає можливості отримати справжні повторні спостережень, розроблені методи, які формують велику кількість так званих "псевдо-вибірок", і на їх основі можна отримати необхідні характеристики шуканого параметра: оцінки математичного очікування, дисперсії, надійного інтервалу. Методи "чисельного ресамплінгу" або, як їх іноді називають "методи генерації повторних вибірок" об'єднують чотири різних підходи, що відрізняються за алгоритмом, але близьких по суті: рандомизация, або перестановочний тест (permutation), бутстрап (bootstrap), метод "складного ножа" (jackknife) і кросперевірка (cross-validation). Ці алгоритми є сучасною альтернативою параметричних методів і бурхливо розвиваються два останніх десятиліття.

Головна причина недостатнього практичного застосування методів чисельного ресамплінгу пояснюється відсутністю необхідного програмного забезпечення. Для реалізації такої можливості можна використовувати програми, які керуються за допомогою меню, такі як SPSS, пакету Statistica, за допомогою обчислювальних інтерактивних середовищ, таких як R, MatLab, SAS, Stats, або самостійної розробки програм з використанням Visual Basic, C++, Delphi та ін.

Метою цієї роботи є розробка та програмна реалізація алгоритмів, що дозволяють розв'язувати специфічні задачі статистичної обробки даних. А саме, на основі досліджених методів чисельного ресамплінгу побудовано алгоритми розв'язання типових задач, які

потім були програмно реалізовані засобами мови R. Зокрема, алгоритм бутстрапа полягає у наступному:

Крок 1. За вибірковими даними $\{x_1, x_2, ..., x_n\}$ відбувається побудова ймовірнісної моделі та оцінюються її параметри.

Крок 2. Випадковим чином з відібраного розподілу з параметрами $\hat{\theta}$ генерується n елементів $x_1^*, x_1^*, ..., x_n^*$ і бутстрап-повторність, отримана таким чином, використовується для розрахунку значень статистики $t^* = T(x^*)$.

Крок 3. Крок 2 виконується B разів, формується бутстрап-розподіл, аналізуються статистики $\{x^{*1}, ..., x^{*j}, ..., x^{*B}\}$.

Постановка задачі

У загальному випадку непараметричні підходи, пов'язані з побудовою емпіричної функції розподілу статистичних характеристик досліджуваної випадкової величини, можливі тільки при наявності повторних спостережень. Однак в екології, як і в економіці, медицині і багатьох інших галузях, можна виконати зріз даних тільки у певному місці і в певний момент часу, а якщо відбирати другу, третю проби і т.д., то це будуть вже дані з іншого місця або ж взяті в інший момент часу [5]. Тому виникає питання: як, маючи лише одну єдину повторність, оцінити значення необхідного нам показника і отримати міру точності цієї оцінки?

Спочатку розглянуто задачу, пов'язану з оцінкою параметрів за вибіркою маси тіла 213 особин ящірки прудкої *Lacerta agilis*. Характер розподілу цього варіаційного ряду далекий від нормального закону, в чому легко переконує квантиль-квантільний графік (рис. 1).

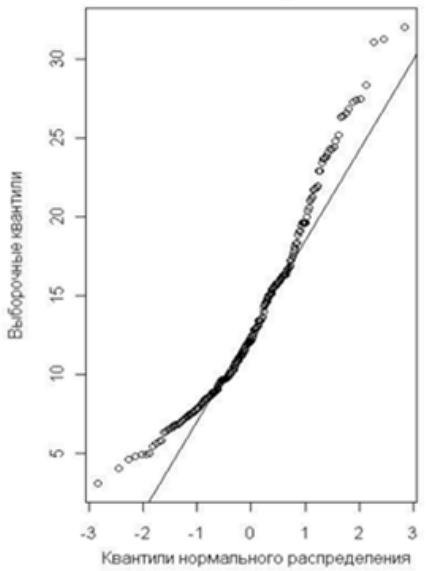


Рис. 1. Графік квантиль-квантильного розподілу вибірки маси тіла ящірки прудкої

Розраховано вибіркові статистики: середня $\bar{x} = 13.65$, стандартне відхилення $s = 6.06$, коефіцієнт варіації $CV = 0.44$, похибка середнього $s_m = 0.415$ і, незважаючи на відсутність необхідних передумов нормальності, двосторонні надійні інтервали середнього з надійністю $\gamma = 95\,\%$:

$$12.83 \le \bar{x} \le 14.46 .$$

Представлено набір можливих методів оцінки інтервальних значень шуканих параметрів з використанням ресамплінгу на прикладі середнього $\bar{x}$ і коефіцієнта варіації

$$CV = \frac{s}{\bar{x}} = \frac{1}{\bar{x}}\left[\sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{(n-1)}\right]^{\frac{1}{2}}.$$

Відзначимо, що похибка коефіцієнта варіації розрахована за наближеною формулою $CV = 0.03$, а двосторонні надійні інтервали з надійністю $\gamma = 95\%$ рівні

$$CI_{0.95} = CV \pm t_{0.025} S_{CV} = 0.444 \pm 2.02 \cdot 0.03 = 0.384 \div 0.504$$

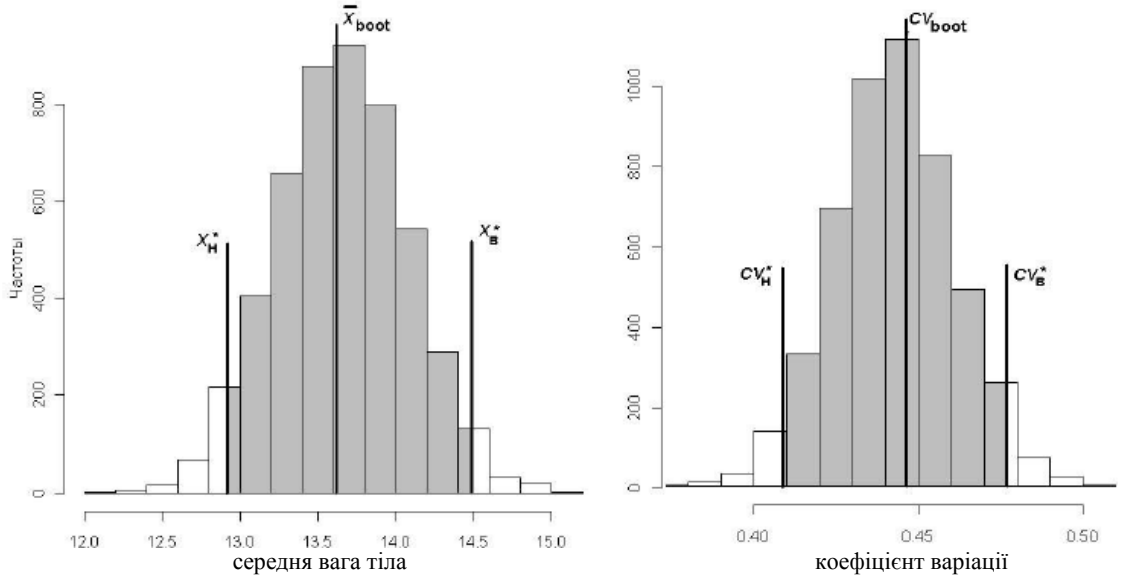
При застосуванні бутстрапа на основі наявних емпіричних даних, з використанням алгоритму заміни з поверненням, згенеровано 5000 псевдовиборок і обчислено для кожної з них оцінки математичного очікування і коефіцієнта варіації. На основі отриманих варіаційних рядів $\{\bar{x}^{*1}, \bar{x}^{*2}, ..., \bar{x}^{*5000}\}$ і $\{CV^{*1}, CV^{*2}, ..., CV^{*5000}\}$ побудовано графіки функції розподілу $\hat{F}$ (рис. 2), знайдено середні значення цих бутстрап-статистик:

- середнього $\bar{x}_{boot} = 13.65$ (зміщення оцінок дорівнює 0.0058);
- коефіцієнта варіації $CV^*_{boot} = 0.4426$ з абсолютним зсувом 0.0014 (0.3 %).

Існують різні теоретичні міркування для отримання інтервальних оцінок параметрів з використанням бутстрапа. Основними умовами реалізації цих методів є симетричність і унімодальність розподілу вихідної вибірки, а також алгоритм бутстрапування, що забезпечує генерацію з цього розподілу випадкових і незалежних повторень. У таблиці 1 наведено оцінки надійних інтервалів та середньої маси тіла ящірок і коефіцієнта варіації, отриманих різними методами.

Таблиця 1. Оцінки надійних інтервалів та середньої маси тіла ящірок і коефіцієнта варіації, отриманих різними методами

Використаний метод оцінки	Надійність, %	Середнє	Коефіцієнт варіації *CV*
1. По вхідній вибірці з використанням *t*-критерія (припущення не дотримується)	95	12.83÷14.46	0.384÷0.504
2. Метод "складного ножа" (jackknife)	95	12.83÷14.46	0.408÷0.481
3. Метод процентилій	95	12.84÷14.48	0.407÷0.479
	90	12.97÷14.34	0.412÷0.473
4. Основні інтервали (basic CI)	95	12.81÷14.45	0.407÷0.478
5. Бутстрап з використанням *t*-критерія	95	12.86÷14.46	0.409÷0.482
6. Інтервали ст'юдентизированого типу	95	12.91÷14.56	0.410÷0.483
7. З корекцією зміщення BCa	95	12.90÷14.49	0.411÷0.484



а) *б*)

Рис. 2. Гістограми частотного розподілу середніх (*а*) та коефіцієнтів варіації (*б*), отримані методом бутстрапування вибірки маси тіла ящірок

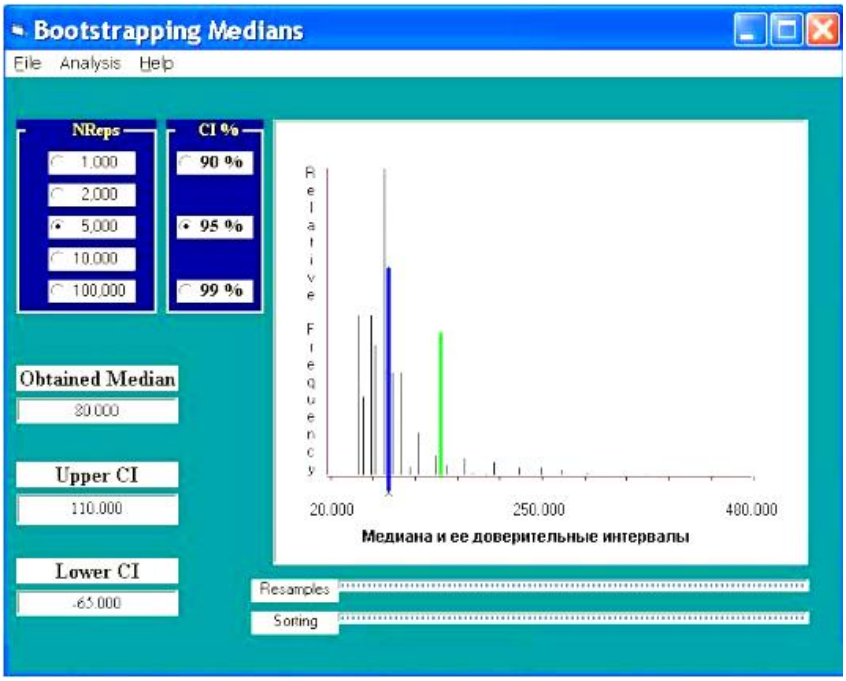


Рис. 3. Оцінка надійних інтервалів медіани чисельності Chironomus salinarius бутстрап-методом

За допомогою бутстрапа можна робити те, що не завжди під силу звичайним параметричним методам. Наприклад, для асиметричних вибірок пропонують використовувати медіану в якості міри оцінки міри положення випадкової величини замість традиційного математичного очікування.

Розглянемо вибірку популяційної щільності (екз/м^2) личинок *Chironomus salinarius* в 43 пробах з різних річок, де цей вид був виявлений (в дужках наведені частоти повторюваних значень): 5 (2); 8; 10 (3); 19; 20 (3); 30; 40; 42; 50 (6); 65 (2); 80 (3); 100 (2); 133; 200; 250; 300; 430; 440; 480; 800; 880; 2400; 3020; 3360; 5200; 6200; 7000; 9000; 19000.

Характер розподілу цього варіаційного ряду відрізняється від нормального закону, що дає підстави припускати, що медіана є однією з найбільш інтерпре-

тованих оцінок. Знайдене вибіркове значення медіани $Me = 80$, що значно менше середнього арифметичного $\bar{x} = 1432$.

Виконано непараметричний бутстрап представленої вибірки і оцінено надійні інтервали медіани чисельності *Chironomus salinarius* (рис. 3).

У загальному випадку, якщо ми маємо результати спостережень, виміряні в шкалах чисельності об'єктів, то кількість зареєстрованих унікальних значень, як правило, невелика. Як би багато не проводилося ітерацій бутстрапа, ми можемо отримати тільки 43 псевдозначень медіан. Оцінка стандартного відхилення по такому частотному розподілу пов'язана з істотною статистичної помилкою, тому нижня межа 95 %-го надійного інтервалу, розрахованого деякими методами, заснованими на стандартній похибці se_{boot}, може зміститися у від'ємну область (рис. 3). Для порівняння в представленому прикладі надійні інтервали, розраховані методом процентилей рівні 50÷200, а методом *BCa* 50÷133.

Описана проблема характерна для багатьох випадків, де вимірюваний показник представлений в балах: класи якості водойм, оцінки знань учнів і т.д. Її рішення може бути здійснено з використанням одного з методів імітації Монте-Карло. Зокрема, генерація будь-якої за обсягом послідовності випадкової величини із заданого розподілу легко здійснюється на основі алгоритму зворотної трансформації.

У більшості екологічних програм пошук базової ймовірнісної моделі для вибіркових даних — складний і неоднозначний процес. Розглянемо підбір теоретичного розподілу для вибірки маси тіла ящірки прудкої Lacerta agilis, статистичні характеристики якої аналізувалися вище (Табл. 1). Ефективним способом перевірки згоди розподілів є побудова квантіль-квантільних QQ-графіків (рис. 1): якщо квантилі нормального розподілу та емпіричні квантилі пропорційні між собою і вибіркові точки строго шикуються на "теоретичній" прямій, то апроксимацію можна вважати вдалою. Однак наскільки статистично значимий розкид експериментальних точок щодо теоретичної прямої QQ? Згенеруємо велике число B випадкових вибірок з нормального розподілу з параметрами, оціненими за вибірковими даними. Ми будемо мати цілий пучок прямих, які незначно відрізняються коефіцієнтами кута нахилу. Якщо провести перпендикулярно через цей пучок січні площини і з'єднати крайні точки кривої, то ми отримаємо "коридор" з двох огинаючих (point-wise envelope), всередині якого розташувалися будь-які зі згенерованих прямих (рис. 4). Можна провести також сімейство огинаючих з різним рівнем довірчої ймовірності, тобто огинаючих, наприклад, 90% бутстрап-прямих.

Так як апроксимація нормальним розподілом виявилася невдалою, було розглянуто два інших базових розподіли, які широко використовують в демографічних дослідженнях — логнормальний та Вейбулла (рис. 5).

Для логнормального розподілу за допомогою бутстрап-процедур побудовано квантиль-квантильний графік з огинаючими, які обмежують прийняті надійні інтервали (рис. 6).

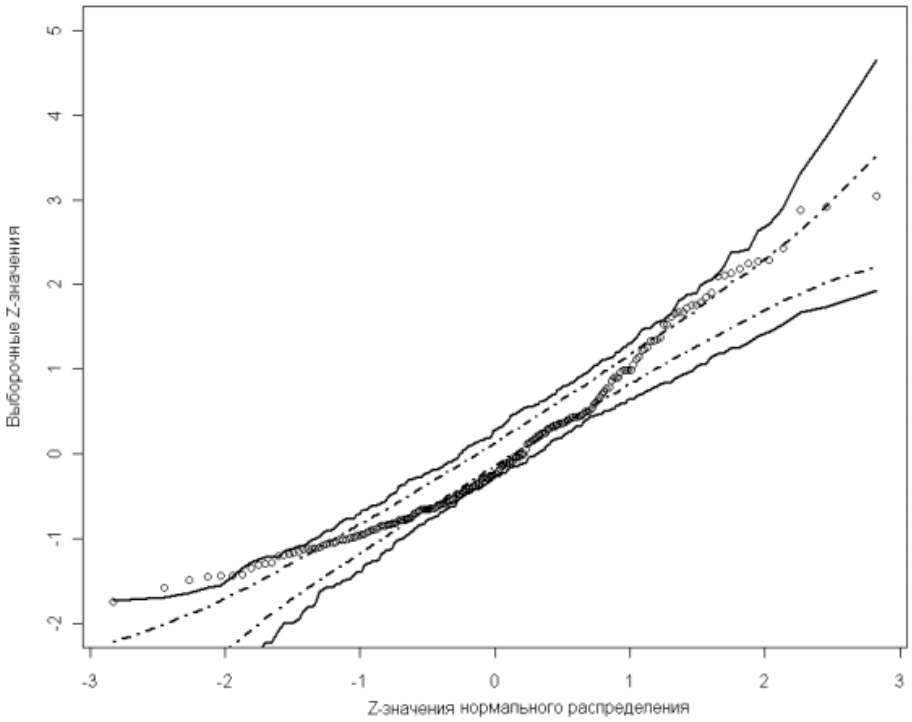
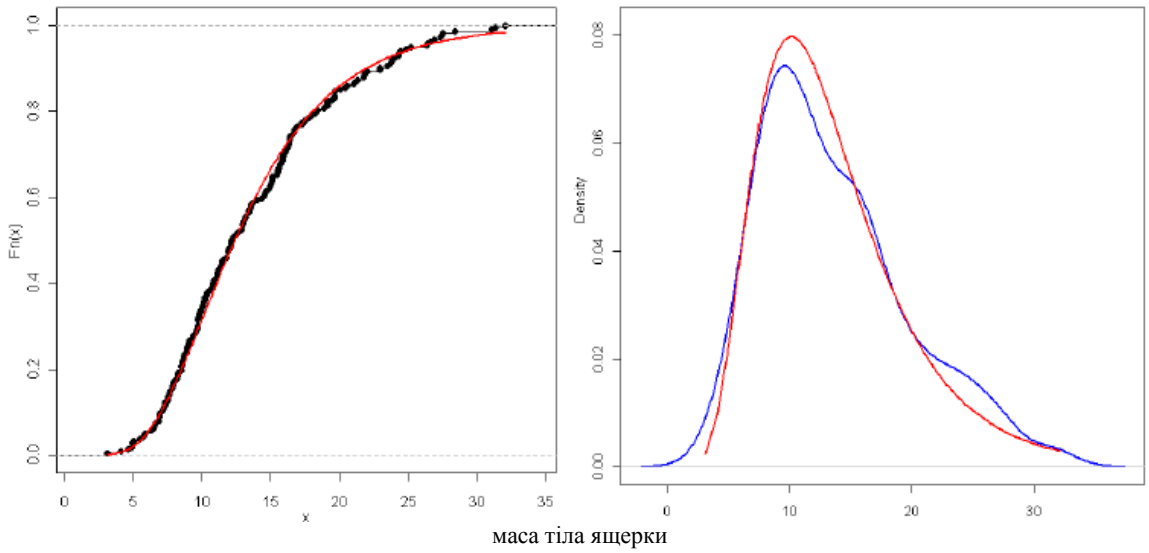


Рис. 4. Графік залежності Z-значень для нормального розподілу і по спостережуваним даними маси тіла ящірки прудкої; показані надійні огинають, отримані після 2000 ітерацій бутстрапа (100 % — суцільна лінія, 90 % — штрих-пунктир)



а) *б*)

Рис. 5. Апроксимація вибіркових даних маси тіл ящірок логнормальним розподілом: емпірична та теоретична кумулятивна функція розподілу ймовірностей (*а*); теоретична функція щільності та функція, що згладжує емпіричний розподіл (*б*)

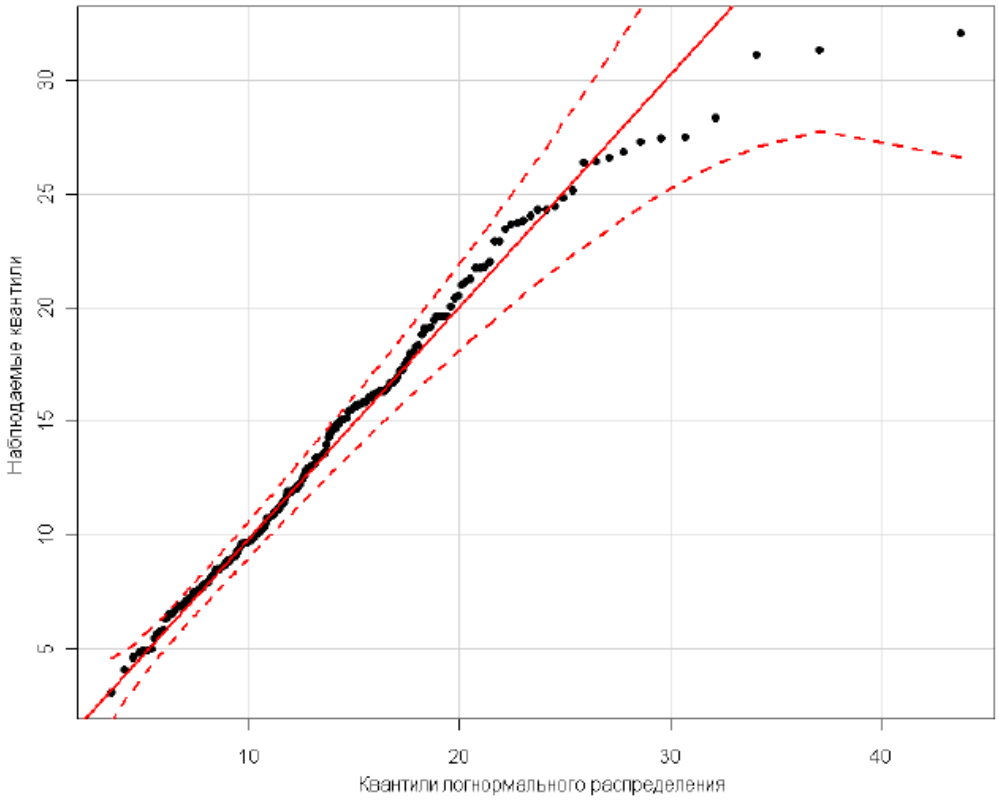


Рис. 6. 95 %-і надійні огинаючі при апроксимації вибіркової маси тіла ящірок логнормальним розподілом

Висновки

При аналізі екологічних даних не достатньо обмежуватись обчисленням точкових оцінок показників, які вивчаються по заданим вибіркам. Слід переконатися у тому, що знайдена величина, наприклад середньої популяційної щільності особин, є точною та незміщеною, а також отримати надійні інтервали. Або якщо порівнювати видовий склад для двох або декількох місцепроживань, то необхідно оцінити ймовірність того, що знайдена збіжність біотопів статистично значимо відрізняється від випадкового розподілу. Тільки так можна обгрунтувати те, що мінливість показників мно-

жини організмів має екологічно закономірний характер. Тут виникає два питання. По-перше, статистичні властивості параметрів можуть бути вивчені тільки при наявності повторних спостережень. Однак, в екології можна виконати зріз даних тільки в певному місці і в певний момент часу, а якщо відбирати другу, третю проби і т. д., то це будуть вже дані з іншого місця або ж взяті в інший момент часу. Отже, як, маючи лише одну єдину повторність, оцінити значення необхідного нам показника і отримати міру точності цієї оцінки? По-друге, оскільки точний вид розподілу оброблюваних даних, зазвичай, невідомий, використовують наближені методи апроксимації передбачуваних властивостей досліджуваної статистики, причому як впливає ступінь цієї наближеності на остаточні висновки, залишається цілком на совісті дослідника. Можна сказати, що вирішення цих проблем може бути здійснено застосуванням методів генерації повторних вибірок (чисельного ресамплінгу), які відносяться до непараметричних методів.

Результати досліджень математичних моделей еколого-інформаційних систем, для яких було застосовані алгоритми непараметричної статистики, дають відповіді на поставленні питання та свідчать про те, що розроблені алгоритми можуть бути застосованими до інших класів прикладних задач, які у своїй математичній постановці зводяться до досліджених.

ЛІТЕРАТУРА

1. Накопія М. Окремі випадки застосування методів чисельного ресамплінгу в прикладних задачах / М.Накопія // Математичне моделювання, Дніпродзержинськ, №2(33) – 2015 – С. 11–13.
2. Efron B. An intrduction to the bootstrap. / B. Efron, R.J.Tibshirani // N. Y.: Chapman &Hall – 1993 – P. 436
3. Manly B.F. Randomization, bootstrap and Monte Carlo methods in biology / B.F. Manly // London:Chapman & Hall – 2007 – P. 445.
4. Анатольев С. Основы бутстрепирования / С. Анатольев // Квантиль – 2007 – №3 – С. 1–12.
5. Шитиков В.К. Рандомизация и бутстреп: статистический анализ в биологии и экологии с ипользованием R / В.К. Шитиков, Г.С.Розенберг // Тольятти: Кассандра – 2013 – С. 314.

пост. 27.02.2017