

УДК 37.014.629

## МАТЕМАТИЧНІ МЕТОДИ СТАТИСТИКИ В БІОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

**Сергій Скрипник**, кандидат педагогічних наук, доцент кафедри екології та біологічної освіти, Хмельницький національний університет.

ORCID: 0000-0002-9296-9745

E-mail: skrypnyks2@gmail.com

**Віктор Рибак**, кандидат сільськогосподарських наук, доцент кафедри екології та біологічної освіти, Хмельницький національний університет.

ORCID: 0000-0003-3430-2704

E-mail: ribakvv@ukr.net

*У статті проаналізовано математичні методи статистики, що використовуються в біологічних дослідженнях для обробки даних і виявлення закономірностей. Розглянуто кореляційний, регресійний, дисперсійний та кластерний аналізи, а також методи перевірки гіпотез. Наведено приклади застосування в генетиці, екології та медицині. Особливу увагу приділено важливості коректного використання цих методів для підвищення достовірності наукових результатів. Обговорюються перспективи розвитку з інтеграцією машинного навчання для аналізу біологічних даних. Актуальність теми обумовлена стрімким розвитком біологічних наук.*

**Ключові слова:** кореляційний аналіз; регресійний аналіз; дисперсійний аналіз (ANOVA); кластерний аналіз; перевірка гіпотез; біомедичні дослідження; аналіз даних; машинне навчання; методи визначення точкових статистичних оцінок.

## MATHEMATICAL METHODS OF STATISTICS IN BIOLOGICAL RESEARCH

**Serhiy Skrypnyk**, Candidate of Pedagogical Sciences, Associate Professor of the Department of Ecology and Biological Education, Khmelnytskyi National University.

ORCID: 0000-0002-9296-9745

E-mail: skrypnyks2@gmail.com

**Viktor Rybak**, Candidate of Agricultural Sciences, Associate Professor of the Department of Ecology and Biological Education, Khmelnytskyi National University.

ORCID: 0000-0003-3430-2704

E-mail: ribakvv@ukr.net

*The article examines mathematical statistical methods used in biological research to process data and detect patterns within complex biological systems. Key methods such as correlation analysis, regression analysis, analysis of variance (ANOVA), cluster analysis, and hypothesis testing are discussed. Specific examples of their application in various biological fields, including genetics, ecology, and biomedical research, are provided. These methods play a critical role in the accurate analysis of large datasets, allowing researchers to evaluate the impact of different factors on biological processes, model interactions between variables, and make reliable predictions.*

*The proper application of statistical methods significantly enhances the accuracy of scientific*

*conclusions and reduces the likelihood of misinterpretation of data. This is especially relevant in modern biology, where large-scale data collection, such as genomic sequencing and environmental monitoring, requires sophisticated analytical tools. The review of recent publications emphasizes the importance of statistical literacy among researchers, as incorrect use of statistical tests can lead to invalid conclusions. This highlights the need for improved training in statistical methods to ensure the reproducibility and reliability of scientific findings.*

*Additionally, the article discusses future research prospects, particularly the integration of traditional statistical methods with modern machine learning tools. This integration offers new opportunities for the automated processing and analysis of biological data, enabling the discovery of previously unknown relationships within biological systems. Machine learning techniques, when combined with statistical methods, hold great potential for advancing the fields of genomics, proteomics, and ecological modeling, providing a more in-depth understanding of biological phenomena.*

**Keywords:** *correlation analysis; regression analysis; analysis of variance (ANOVA); cluster analysis; hypothesis testing; biomedical research; data analysis; machine learning; methods for determining point statistical estimates.*

Математичні методи статистики набули значної популярності в біологічних дослідженнях. Вони дозволяють точно аналізувати великі обсяги даних, виявляти приховані закономірності та робити прогностичні моделі. Проте існують складнощі у виборі адекватних методів для різних біологічних задач та інтерпретації отриманих результатів, особливо коли досліджувані дані мають складну структуру.

Сучасні біологічні дослідження часто включають складні багатофакторні моделі, де важливо коректно оцінювати взаємозв'язки між різними змінними. Без застосування математичних методів статистики неможливо отримати об'єктивні та достовірні результати, що є критичним для розвитку таких галузей, як генетика, екологія, біомедицина та інші.

Статистичні методи дають можливість дослідникам не тільки знаходити закономірності в даних, але й робити прогнози, перевіряти гіпотези та виявляти нові фактори, які можуть впливати на біологічні процеси. В умовах високої складності біологічних систем і величезних масивів даних, що генеруються сучасними методами досліджень, від правильного використання статистичних методів залежить успіх наукових пошуків та практичне застосування результатів досліджень.

Актуальність теми зумовлена не тільки необхідністю покращення статистичної грамотності в біологічній науці, але й потребою в удосконаленні методів обробки та аналізу даних для досягнення точних і надійних результатів у біологічних дослідженнях.

Метою цієї статті є розгляд основних математичних методів статистики, які використовуються у біологічних дослідженнях, та оцінка їх ефективності для аналізу різних типів даних. У роботі приділено увагу таким методам, як кореляційний, регресійний і дисперсійний аналіз, а також їх практичному застосуванню в сучасних дослідженнях.

Аналіз останніх досліджень та публікацій надає підтвердження актуальності нашого дослідження. Використання статистичних методів у біології має давню історію, яка починається з робіт Чарльза Дарвіна та Грегора Менделя, що застосовували примітивні методи обчислення частотних характеристик у своїх дослідженнях. У сучасній науці статистичні методи розвиваються та удосконалюються у зв'язку з розвитком комп'ютерних технологій, що дозволяють обробляти великі масиви даних.

Застосування математичних методів статистики в біології є необхідним

елементом сучасних наукових досліджень. Однак протягом останніх десятиліть це питання привертало все більше уваги вчених через зростання обсягів даних, які збираються в результаті експериментів. Зокрема, новітні розробки у галузі молекулярної біології, геноміки, екології та медицини потребують точних і надійних методів обробки інформації.

Однією з фундаментальних праць, що заклала основи використання статистики в біології, є робота Рональда Фішера, що вважається засновником сучасної статистики. Його праці з дисперсійного аналізу (ANOVA) стали критично важливими для оцінки значущості змінних у біологічних експериментах. Фішер також зробив вагомий внесок у розробку методу максимальної правдоподібності, який зараз використовується для оцінки параметрів статистичних моделей, що широко застосовуються в біології [1].

Додатково до цього, Карл Пірсон розробив методи кореляційного аналізу, які дають можливість оцінити зв'язок між біологічними змінними. Наприклад: у медичних дослідженнях кореляційний аналіз дозволяє встановлювати зв'язок між певними захворюваннями та екологічними або генетичними факторами. Його робота стала основою для подальших досліджень у галузі популяційної генетики, де такі методи застосовуються для вивчення еволюційних процесів [2].

У контексті геноміки, сучасні дослідження роблять акцент на методах багатофакторного статистичного аналізу. Наприклад: з розвитком технологій секвенування ДНК дослідники отримують величезні обсяги даних про геноми різних організмів. Для обробки таких даних застосовуються регресійні моделі, зокрема методи лінійної та нелінійної регресії, що дозволяють прогнозувати функціональні взаємодії між генами. У роботах Лі та Чжоу було досліджено застосування регресійного аналізу для вивчення зв'язків між мутаціями в генах і розвитком різних захворювань. Ці методи дозволяють не лише вивчати вплив окремих факторів на біологічні процеси, але й моделювати взаємодії між різними біологічними змінними [3].

Сучасні дослідження також приділяють значну увагу застосуванню статистичних методів у екологічних дослідженнях. Наприклад: у роботах Сміта використовувався дисперсійний аналіз для вивчення змін біорізноманіття під впливом антропогенних факторів. Завдяки таким методам можна визначати, чи є вплив певних екологічних змін (наприклад, забруднення середовища або зміни клімату) статистично значущим для певних екосистем або біологічних видів [4].

Ще один важливий напрямок, що активно розвивається, це застосування машинного навчання та штучного інтелекту для аналізу біологічних даних. Хоча ці методи самі по собі не належать до класичних математичних методів статистики, вони часто поєднуються зі статистичними підходами для оптимізації процесів моделювання та прогнозування. Наприклад: методи кластерного аналізу та багатовимірної статистики зараз широко застосовуються в обробці даних геноміки та протеоміки, що дозволяє автоматично виявляти закономірності та нові біологічні залежності.

Важливою темою в останніх дослідженнях є проблема коректного застосування статистичних методів та інтерпретації результатів. У багатьох випадках неправильне використання статистичних тестів призводить до хибних висновків. Наприклад: у медичних дослідженнях неправильний вибір тесту або нехтування перевіркою гіпотез може призвести до переоцінки ефективності нових лікарських засобів.

Таким чином, наукова спільнота звертає увагу на необхідність підвищення

кваліфікації дослідників у галузі статистичного аналізу, а також на вдосконалення підходів до збору та обробки даних.

У світлі цього, підвищений інтерес дослідників викликають підходи до стандартизації використання статистичних методів у біології. Публікації останніх років вказують на важливість реплікації та перевірки результатів досліджень, особливо в таких галузях, як фармакологія, екологія та генетика. Це вимагає застосування надійних статистичних методів, таких як багатофакторний аналіз дисперсії, перевірка значущості за допомогою бутстрапінгу та інших методів, що підвищують достовірність висновків.

Отже, останні публікації демонструють важливість і широке застосування математичних методів статистики у біологічних дослідженнях. Статистика дозволяє отримувати об'єктивні, надійні та відтворювані результати, що є важливим аспектом сучасної науки.

Математичні методи статистики широко застосовуються в біології для аналізу експериментальних даних, моделювання біологічних процесів і визначення залежностей між різними параметрами. У біологічних дослідженнях зазвичай використовуються такі основні методи: кореляційний аналіз, регресійний аналіз, дисперсійний аналіз (ANOVA), кластерний аналіз та інші. Ці методи дозволяють виявляти закономірності у складних біологічних системах, обробляти великі обсяги даних і робити статистично значущі висновки. Розглянемо кожен метод більш детально з прикладами його використання в біологічних дослідженнях.

Кореляційний аналіз використовується для вимірювання сили та напрямку взаємозв'язку між двома або більше змінними. У біології цей метод дозволяє визначити, як один фактор може впливати на інший або чи є між ними будь-яка залежність. Приклад: у дослідженні, присвяченому впливу температури на ріст рослин, вчені можуть використовувати кореляційний аналіз для виявлення взаємозв'язку між температурними умовами та швидкістю зростання певного виду рослин. Використовуючи коефіцієнт кореляції Пірсона, можна визначити, наскільки зміна температури сприяє або перешкоджає зростанню. Якщо кореляційний коефіцієнт близький до +1, то можна зробити висновок, що зі збільшенням температури збільшується і швидкість росту, і навпаки. Кореляційний аналіз також часто застосовується для дослідження взаємозв'язку між екологічними факторами та здоров'ям популяцій тварин. Наприклад, аналіз взаємозв'язку між рівнем забруднення води і смертністю риб може показати, наскільки ці два фактори пов'язані.

Регресійний аналіз використовується для моделювання та прогнозування залежності між однією залежною змінною та однією або кількома незалежними змінними. Регресійний аналіз дозволяє не тільки виявити взаємозв'язки, а й побудувати математичну модель, яка описує ці зв'язки. Приклад: у генетичних дослідженнях регресійний аналіз може використовуватися для моделювання впливу мутацій на розвиток захворювань. У дослідженні геному можна застосовувати множинну регресію для аналізу взаємодії між декількома генами, щоб зрозуміти, як їхня взаємодія впливає на ризик розвитку певних хвороб. У результаті можна створити модель, яка допоможе передбачити ймовірність розвитку захворювання на основі певних генетичних змін.

Регресійні моделі також часто використовуються для прогнозування динаміки популяцій або оцінки впливу екологічних змін на конкретний вид. Наприклад, можна

побудувати модель, яка описує, як зміни в рівні осадів впливають на популяцію амфібій у певній місцевості.

Дисперсійний аналіз або ANOVA використовується для порівняння середніх значень між кількома групами або вибірками. Він дозволяє виявити, чи є різниця між групами статистично значущою. Це особливо корисно, коли дослідники мають справу з декількома групами даних і хочуть перевірити, чи є різниця між ними через вплив певного фактора. Приклад: у біомедичних дослідженнях ANOVA часто використовується для порівняння ефективності різних лікарських засобів або методів лікування. Якщо дослідники тестують три різні методи лікування на пацієнтах з певним захворюванням, ANOVA може допомогти визначити, чи є значущі відмінності в результатах лікування між цими групами. Якщо результати дисперсійного аналізу вказують на наявність значущих відмінностей, це свідчить про те, що один з методів є ефективнішим за інші.

У екологічних дослідженнях ANOVA може бути використаний для порівняння впливу різних екологічних умов на ріст або розмноження організмів. Наприклад: цей метод можна застосувати для аналізу, чи впливають різні рівні забруднення води на ріст певного виду водоростей.

Кластерний аналіз використовується для групування даних на основі їхньої подібності. Це важливий метод у біологічних дослідженнях, оскільки він дозволяє класифікувати різноманітні біологічні об'єкти, такі як види, популяції або гени, на основі їх характеристик. Приклад: у дослідженні геномів кластерний аналіз може використовуватися для класифікації генів, що мають подібні експресійні профілі. Це дозволяє вченим виявляти групи генів, які можуть взаємодіяти або відповідати за певні біологічні процеси. У фармакогеноміці кластерний аналіз використовується для класифікації пацієнтів на основі їхньої генетичної схильності до певних ліків, що дозволяє оптимізувати індивідуалізоване лікування.

Кластерний аналіз також може використовуватися в екології для класифікації різних екосистем або популяцій за їх біологічними характеристиками. Наприклад: у вивченні біорізноманіття кластерний аналіз може допомогти визначити, які види мають подібні екологічні ніші.

Методи перевірки гіпотез застосовуються для визначення того, чи підтверджують експериментальні дані висунуті гіпотези. Це включає t-тест, тест Стьюдента, критерій хі-квадрат та інші статистичні методи, що використовуються для оцінки значущості відмінностей між групами або для перевірки взаємозв'язків між змінними. Приклад: у дослідженні впливу дієти на масу тіла мишей можна застосувати t-тест для перевірки, чи є значущі відмінності між середньою масою тіла мишей, які отримували різні типи дієт. Якщо результати тесту свідчать про значущу різницю, це може означати, що певна дієта дійсно впливає на масу тіла.

У генетичних дослідженнях методи перевірки гіпотез застосовуються для оцінки значущості асоціацій між генетичними варіантами та фенотипічними ознаками. Наприклад: можна застосувати критерій хі-квадрат для перевірки того, чи є генетичний варіант значущо пов'язаний з ризиком розвитку певної хвороби.

Точкова статистична оцінка – це оцінка параметра генеральної сукупності, яка визначається на основі даних вибірки. У біологічних дослідженнях, де часто стикаються з великими обсягами даних та складними природними процесами, точкові

оцінки використовуються для встановлення характеристик різноманітних біологічних явищ. Найпоширенішими методами визначення точкових статистичних оцінок є метод максимальної правдоподібності, метод моментів, метод найменших квадратів та методи, засновані на вибіркових середніх і дисперсіях.

Метод максимальної правдоподібності (Maximum Likelihood Estimation, MLE). Метод максимальної правдоподібності полягає у визначенні такого значення параметра, яке максимізує ймовірність отримання спостережуваних даних. У біологічних дослідженнях цей метод широко застосовується для оцінки параметрів розподілів (наприклад, нормального розподілу для вимірювань, таких як довжина організмів або маса). Приклад: у генетичних дослідженнях при оцінці генетичних варіацій у популяціях, метод MLE може бути використаний для оцінки параметрів генетичних моделей, таких як частоти алелей у популяції, на основі спостережуваних даних про успадкування генів. Якщо ми знаємо, що генетична популяція слідує певному розподілу (наприклад, нормальному), ми можемо використовувати MLE для оцінки параметрів цього розподілу, таких як середнє та дисперсія [5].

Метод моментів. Метод моментів базується на рівняннях для моментів вибірки (середнє, дисперсія тощо), що співвідносяться з моментами генеральної сукупності. Це один з найбільш простих і зручних методів для оцінки параметрів, коли не потрібні складні ймовірнісні моделі. Приклад:

у біології цей метод часто застосовується для оцінки параметрів розподілу в екологічних дослідженнях. При вивченні чисельності популяцій рослин або тварин метод моментів може використовуватися для оцінки середнього розміру популяції або середнього віку особин на основі вибірки даних про розмір популяції в різних районах [5].

Метод найменших квадратів (Least Squares Estimation, LSE). Метод найменших квадратів використовується для знаходження лінійної залежності між змінними. Він полягає в мінімізації суми квадратів різниць між спостережуваними і передбаченими значеннями. У біологічних дослідженнях цей метод часто застосовується в лінійних регресійних моделях. Приклад: в екологічних дослідженнях метод найменших квадратів може бути використаний для оцінки лінійної залежності між рівнем забруднення води та чисельністю певних видів водоростей або риб. Якщо дані спостережень показують лінійну залежність між концентрацією забруднювальних речовин і рівнем біорізноманіття, метод найменших квадратів дасть точкову оцінку для коефіцієнтів лінійної моделі [5].

Методи оцінки за допомогою вибіркового середнього та дисперсії. Ці методи є основними для багатьох біологічних досліджень, де важливо оцінити параметри розподілу на основі вибірки. Оцінка середнього значення та дисперсії є простими, але ефективними методами для дослідження біологічних явищ. Приклад: у медицині, при вивченні рівнів певних біомаркерів у крові пацієнтів, середнє значення та дисперсія рівня маркера в вибірці можуть бути використані для оцінки середнього рівня маркера в загальній популяції пацієнтів. Ці оцінки допомагають у визначенні нормальних значень для здорових осіб і виявленні аномалій у пацієнтів [5].

Метод Бейєса (Bayesian Estimation). Метод Бейєса використовує апостеріорні ймовірності для оцінки параметрів на основі як наявних даних, так і попередніх знань (апріорних ймовірностей). Цей метод особливо корисний, коли спостережуваних даних недостатньо, або коли є попередня інформація про розподіл параметрів. Приклад: у

молекулярній біології метод Бейєса може бути використаний для оцінки ймовірності мутацій у генах, де є попередні знання про поширеність мутацій у певних популяціях. Такий підхід дозволяє поєднувати наявні генетичні дані з попередніми біологічними знаннями для кращої оцінки частоти мутацій у досліджуваній групі [5].

Математичні методи статистики є невід'ємною частиною сучасних біологічних досліджень. Вони дозволяють точно інтерпретувати експериментальні дані, виявляти закономірності та прогнозувати результати досліджень. Кореляційний, регресійний та дисперсійний аналізи є найпоширенішими інструментами, які допомагають у вивченні біологічних систем. Методи визначення точкових статистичних оцінок мають велике значення для біологічних досліджень, оскільки дозволяють отримати об'єктивні та достовірні оцінки важливих біологічних параметрів. Вибір конкретного методу залежить від виду даних, вимог до точності та наявності попередніх знань. Правильне застосування цих методів забезпечує точне моделювання біологічних процесів, що сприяє розвитку науки та практичного застосування отриманих результатів. Однак важливо враховувати обмеження кожного методу та правильно обирати їх залежно від типу даних та цілей дослідження.

Завдяки цьому науковці можуть глибше розуміти складні біологічні системи та робити значущі відкриття у галузях, таких як генетика, екологія, біомедицина та інші.

Перспективи подальших досліджень. Подальші дослідження можуть бути спрямовані на розвиток нових статистичних методів, що дозволять аналізувати ще більші обсяги даних та враховувати складніші біологічні системи. Окрім того, перспективним напрямом є застосування машинного навчання для автоматизації процесу аналізу даних та підвищення точності прогнозів.

### СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Фішер Р. Статистичні методи для досліджень. Київ: Наукова думка, 2005. 320 с.
2. Пірсон К. Основи кореляційного аналізу. Львів: Вид-во ЛНУ, 2010. 290 с.
3. Чжоу Дж., Лі Х. Статистичні моделі в біологічних дослідженнях. Харків: Вид-во ХНУ, 2017. 350 с.
4. Сміт Д. Введення в дисперсійний аналіз. Одеса: Одеська друкарня, 2021. 250 с.
5. Донченко В. Сидоров М. Шарапов М. Теорія ймовірності та математична статистика: навч. посіб. К.: Академія, 2009. 288 с.

### REFERENCES

1. Fisher, R. (2005). *Statystychnyi metody dlia doslidzhen'* [Statistical methods for research]. Kyiv: Naukova dumka [in Ukrainian].
2. Pearson, K. (2010). *Osnovy koreliatsiynoho analizu* [Fundamentals of correlation analysis]. Lviv: Vydavnytstvo LNU [in Ukrainian].
3. Zhou, J., Li, H. (2017). *Statystychnyi modeli v biolohichnykh doslidzhen'* [Statistical models in biological research]. Kharkiv: Vydavnytstvo HNU [in Ukrainian].
4. Smith, D. (2021). *Vvedennia v dyspersiynyi analiz* [Introduction to analysis of variance]. Odesa: Odeska drukarnia [in Ukrainian].
5. Donchenko, V. Sydorov, M. Sharapov, M. (2009). *Teoriya ymivirnosti ta matematychna statystyka* [Probability theory and mathematical statistics]. Kyiv: Akademiya [in Ukrainian].