Лекція № 4

Аналіз структури екологічних систем. Методи математичної статистики і теорії ймовірності у моделюванні та прогнозуванні стану довкілля

План

1. Основні поняття математичної статистики і теорії ймовірності
2. Аналіз структури та дослідження взаємозв'язків у екологічних системах

1. Основні поняття математичної статистики і теорії ймовірності

При дослідженні екосистем чи техноекосистем з метою подальшого моделювання та прогнозування певних їх параметрів наважливішими аспектами є аналіз структури системи та аналіз її динаміки. Останній буде більш детально висвітлений в наступному параграфі. Щодо структури, то для її вивчення найчастіше застосовуються методи еколого-стохастичного моделювання, що грунтуються на математичному апараті математичної статистики і теорії ймовірності. В даному посібнику не має можливості повно та грунтовно викласти основи цих розділів математики, тому надалі ми зупинемось лише на висвітлені основних понять та найпоширеніших методів.

Величезний потік інформації про стан навколишнього природного середовища, який сьогодні приходиться обробляти екологічній науці, широке та інтенсивне запровадження кількісних методів, що диктується нагальною необхідністю вирішувати важливі прикладні задачі охорони природи, вимагають від сучасного спеціаліста-еколога оволодіння і застосування математико-статистичних методів. Статистичний аналіз суттєво розширює в наш час область застосування. Багата база емпірично отриманих даних про стан навколишнього природного середовища, результати фонового екологічного моніторингу дає можливість для більш точного кількісного та більш глибокого якісного вивчення явищ і об'єктів. Разом з тим подальші потреби розвитку екологічної науки та практики ставлять задачі дослідження складних явищ, сукупностей взаємопов'язаних об'єктів, систем. Все це призводить до необхідності виявлення і вивчення кількісних і якісних закономірностей, багато із якиїх мають статистичний характер.

Статистика – це галузь знань чи практичної діяльності, спрямована на збирання, групування, обробку та інтерпретацію даних. В перекладі із латині "*statistica*" – сума знань про державу. Вперше в сучасному розумінні термін був запроваджений Г. Ахенвалем (1719-1772). Предметом математичної статистики є статистична сукупність.

Статистична сукупність – це множина індивідуально відмінних об'єктів, що володіють спільними властивостями. Статистична сукупність може бути утворена за одним чи декількома ознаками. Кожний елемент цієї сукупності характеризується певним значенням цієї ознаки (властивості), що і є об'єктом вивчення. Окреме значення групувальної озаки називається

варіантою. Число, що показує скільки разів дана варіанта спостерігалась в досліджуваній сукупності називається ***частотою***. Якщо розділити частоти на загальну величину статистичної сукупності (кількість варіант у ряді), то отримаємо – ***частості***. Вони виражаються в долях одиниці, а їх сума дорівнює 1.

Ряд статистичних даних, який отримано в результаті їх зведення і групування за певною змінною кількісною чи якісною ознакою, називається ***рядом розподілу.*** Розрізняють ряди розподілу ***атрибутивні*** – утворені за змінними якісними ознаками та ***варіаційні*** – утворені за кількісними ознаками. Останні застосовуються більш широко. Для того, щоб отримати варіаційний ряд, слід розмістити варіанти у зростаючому чи спадному порядку і вказати відносно кожної її частоту. Тому варіаційний ряд найчастіше представляють у вигляді таблиці з двума колонками, в одній із яких – значення варіант упорядкованої сукупності, а в іншій – частоти (табл. 1). Варіаційні ряди, в свою чергу, поділяються на ***дискретні*** – побудовані за перервними значеннями варіант та ***інтервальні*** – побудовані, відповідно, за неперервними значеннями варіант. Останні можуть мати ***однакові*** або ***неоднакові*** інтервали.

Таблиця 1. *Приклад представлення дискретного варіаційного ряду*

pH, x	6,8	6,9	7,0	7,1	7,2	7,3	7,4	7,5
Частота, f	2	1	1	7	11	5	9	4
Накопичена частота, S_f	2	3	4	11	22	27	36	40

В таблиці 1 представлені результати визначення кислотності 40 проб грунту, відібраних в межах індустріальних зон м. Луцька (за [ ]). Перший і другий рядок таблиці являють собою стандартний варіаційний ряд, в якому навпроти кожного окремого значення pH наведено його частоту (f) і накопичену частоту (S_f), яку ми дещо пізніше використаємо при побудові графіків розподілу.

Таблиця 2. *Приклад представлення інтервального варіаційного ряду*

Межі інтервалів, pH	$\leq 7{,}0$	7,1-7,4	$\geq 7{,}5$
Частота, f	4	32	4

В таблиці 2 варіаційний ряд з попередньої таблиці розбитий на 3 інтервали: з кислотністю pH до 7,0; 7,1-7,4 і більше або рівно 7,5.

В інтервальному варіаційному ряду розрізняють верхню і нижню межу інтервалу (класу). В кожний інтервал включають варіанти, числові значення яких рівні або відповідно більші (менші) нижньої (верхньої) межі класу, згідно умов включення, які можуть бути строгими – з використанням знака =, або нестрогими – з використанням знаків $\leq$ або $\geq$. Інтервальний варіаційний ряд можна представити не лише у вигляді таблиці, але й у вигляді графіка. Графік будується в прямокутній системі координат. На одній з осей (найчастіше – осі

абсцис) відкладають інтервали або серединні значення класів, а по іншій осі – частоти або частості. В залежності від особливостей побудови (рис. 1) даний графік може бути представлений у вигляді: гістограми (для інтервального ряду), полігону (для дискретного ряду), кумуляти або огіви (для обох рядів).

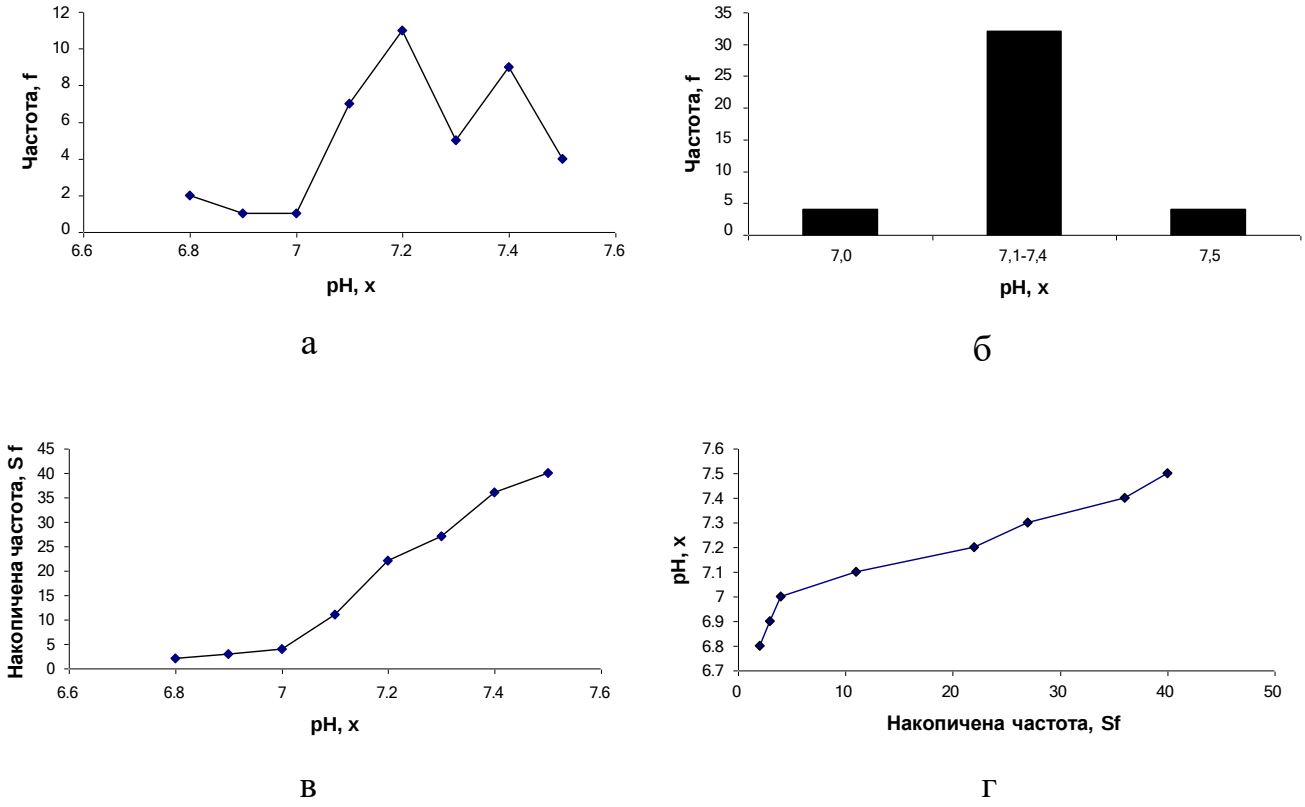


Рис. 1. *Приклади графічного відображення варіаційного ряду, наведеного в таблицях 1,2*:
А – полігон, Б – гістограма, В – кумулята, Г – огіва.

Для того, щоб побудувати ***полігон розподілу***, слід на осі абсцис відмічати точки, що відповідають серединам класів, на осі ординат – частоти (частості), а потім сполучити отримані точки прямими відрізками. Отримана ліміна лінія і є ***полігоном розподілу*** (рис. 1.а).

Гістограму або ***стовпчасту діаграму*** можна отримати, якщо зобразити інтервальний ряд сукупністю прямокутників із спільною основою. Ширина кожного пряокутникука пропорційна величині класу, а його висота – частоті або частості класу. Гістограма для таблиці 2 зображена на рис. 1.б.

Кумулята (***кумулятивна крива***) – зображення в прямокутній системі координат варіаційного ряду з накопиченими частотами. Для її побудови по осі абсцис відкладається значення ознаки, а по осі ординат – накопичені частоти або частості. Потім отримані точки сполучають ламаною лінією, яка і є ***кумулятою*** (рис. 1.в).

Якщо, навпаки, по осі ординат відкладати значення ознаки, а по осі абсцис – накопичені частоти або частості, а потім отримані точки з'єднати

ламаною лінією, отримаємо ***огіву*** (рис. 1.г).

Формуваня вибірки. Для того, щоб отримати виключну інформацію про статистичну сукупність, потрібно повністю врахувати її склад. На практиці це виявляється надзвичайно складно, витратно, а іноді й просто неможливо. Наприклад, провести детальні геохімічні чи гідрохімічні дослідження великих територій та побудувати поле концентрації забруднюючих речовин, яке складається із нескінченної кількості точок. Тому аналізу піддається, як правило, не вся сукупність, а лише деяка її частина, висновки ж пізніше поширюються на всю сукупність. Сукупність, із якої відбираються варіанти для подальшого статистичного вивчення називається ***генеральною сукупністю***. Відібрані із генеральної сукупності варіанти утворюють ***вибірку*** або ***вибіркову сукупність***. Сутність вибіркового методу полягає в тому, що за властивостями вибірки (частини) можна судити про характеристики генеральної сукупності (цілого). Для того, щоб дане судження було правильним, необхідно, щоб вибірка була достатньо ***репрезентативною***. З поняттям репрезентативності тісно пов'язане поняття ***похибки вибірки*** – розбіжності між показниками генеральної та вибіркової сукупності. Розрізняють ***систематичну*** і ***випадкову*** похибку вибірки. Перша виникає внаслідок порушення правил випадкового відбору елементів вибіркової сукупності, а інша – в результаті недостатньо рівномірного представлення у вибірковій сукупності різних категорій елментів генеральної сукупності. Якщо систематичних похибок можна уникнути, то випадкових похибок уникнути не вдається. Їх величина залежить від обсягу вибірки, способу її формування, характеру коливання (варіації) ознаки в генеральній сукупності.

Основними перевагами вибіркового методу є [фещур]:

- практичність;
- зниження реєстрації помилок;
- проведення статистичних досліджень, пов'язаних із знищенням зразків;
- зниження затрат на проведення дослідження та обробку його результатів;
- швидкість та оперативність дослідження.

Найчастіше розрізняють такі методи формування вибіркових сукупностей []:

- простий випадковий відбір;
- систематичний (механічний) відбір;
- типовий (районований) відбір;
- серійний відбір.

Простий випадковий відбір здійснюється шляхом вибору варіант без попереднього розчленування генеральної сукпності на окремі групи (класи). Цей спосіб забезпечує хороший результат, якщо між варіантами генеральної сукупності немає різких відмінностей.

При ***систематичному відборі*** досліджують одиниці сукупності, які розташовані на однаковій відстані та у певній послідовності серед

впорядкованої генеральної сукупності. При цьому задається початок відбору, крок відліку та обсяг вибірки. В геоекології систематична вибірка використовується при зчитуванні інформації із спеціалізованих карт: у вибірку заноситься значення картованого параметра у вузлових (систематичних) точках, наприклад, у вузлах координатної сітки.

При ***типовій вибірці*** формування вибіркової сукупності відбувається на основі попередньої структуризації генеральної сукупності і незалежного відбору елементів із кожної групи (типу). В геоекології дана вибірка найчастіше називається районованою. Для отримання районованої вибірки за картографічним матеріалом потрібно досліджувану території поділити на складові частини, потім одиниці вибірки відбираються по кожному району окремо. Число одиниць, що відбираються в вибірку по кожному району приймається пропорційним його площі.

При ***серійній вибірці*** випадковим чином вибирають групи (серії) одиниць із генеральної сукупності, які повністю досліджуються і їх статистичні властивості екстраполюються на всю сукупність.

Використання того чи іншого способу формування вибіркової сукупності залежить від можливостей спостереження та його мети. Найдостовірніші результати отримують, комбінуючи різні способи формування вибірки.

Попередня обробка даних. Провівши вибір із генеральної сукупності, виміряють одну чи декілька характеристик елементів. Якщо характеристика якісна, то проводять якимось чином її класифікацію і приписують кожній окремій класифікаційній групі (класу) якесь ціле число, що називається ***рангом***. З допомогою рангів класи впорядковуються і надалі стають придатними для кількісного аналізу. Така процедура називається ***ранжуванням***.

Аналіз варіаційного ряду розподілу полягає у виявленні закономірностей зміни частот залежно від зміни кількісної ознаки, яка покладена в основу групування. При аналізі варіаційних рядів найбільш важливими є такі групи показників:

- характеристики центру розподілу;
- характеристики розміру варіації;
- характеристики форми розподілу.

Центром розподілу називається таке значення змінної ознаки, навколо якого групуються інші варіанти. До характеристик центру розподілу відносяться середня величина, мода і медіана.

Середня величина – це інформативна міра “центрального положення” досліджуваної змінної величини. Середнє значення показує типову, характерну величину ознаки, віднесену до одиниці статистичної сукупності. В середній величині нівелюються індивідуальні відмінності одиниць сукупності, які зумовлені дією випадкових факторів, і виявляються загальні закономірності. Виділяються наступні форми середньої величини: середня арифметична, середня гармонійна, середня геометрична, середня квадратична.

Середня арифметична – найпоширеніша форма середньої величини. Її використовують для характеристики рядів розподілу, сума окремих значень ознаки в яких утворює загальний обсяг ознаки:

$$x_a = \frac{\sum x}{n}, \tag{1}$$

де $\bar{x}_a$ – середня арифметична, x – значення ознаки, n – кількість варіант ряду.

Середня гармонійна є величиною, яка обернена до середньої арифметичної з обернених значень ознаки:

$$\bar{x}_h = \frac{n}{\sum \frac{1}{x}}, \tag{2}$$

де $\bar{x}_h$ – середня гармонійна. Середня гармонійна використовуються коли не задана чисельність сукупності і варіанти зважуються за значеннями ваг.

Середня геометрична використовується при аналізі рядів динаміки для розрахунку середніх коефіцієнтів (темпів) зміни. Її ще іноді називають динамічною середньою. Розраховується за формулою:

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}, \tag{3}$$

де $\bar{x}_g$ – середня геометрична.

Середню квадратичну використовують при розрахунках абсолютних і відносних показників варіації ознаки:

$$\bar{x}_q = \sqrt{\frac{\sum x^2}{n}}, \tag{4}$$

де $\bar{x}_q$ – середня квадратична.

Кожна із цих середніх величин може розраховуватись також в зваженому вигляді. ***Зважування середньої величини*** полягає у домножуванні чисельника та знаменника розрахункової формули середньої величини на частоту реалізації ознаки f, а для середньої геометричної – у піднесенні x під дробом степеня Σf до степеня f_i, де і ϵ (1; n).

Окрім середніх величин використовують ще поняття моди та медіани. ***Модою*** називають таке значення ознаки, яке найчастіше зустрічається у сукупності. Для дискретного варіаційного ряду модою є варіанта, що має найбільшу частоту або частість. Для інтервального варіаційного ряду з однаковою шириною інтервалів моду розраховують за формулою:

$$M_0 = \underline{x}_{M_0} + h \cdot \frac{f_{M_0} - f_{M_0-1}}{(f_{M_0} - f_{M_0-1}) + (f_{M_0} - f_{M_0+1})}, \tag{5}$$

де M_0 – мода; $\underline{x}_{M_0}$ – нижня межа модального інтервалу; h – довжина інтервалу; f_{M^0}, f_{M^0-1}, f_{M^0+1} – частоти модального, передмодального і післямодального інтервалів, причому ***модальним*** називається інтервал, що характеризується найвищю частотою ознак варіант, які входять до нього.

Якщо довжина інтервалів варіаційного ряду неоднакова, то для розрахунку моди потрібно звести ряд до інтервального з однаковими інтервалами. Варіаційний ряд може мати одну чи декілька мод. Наявність декількох мод свідчить про неоднорідність сукупності, тобто про об'єднання в одній сукупності різноякісних одиниць. найлегше визначити величину моди графічно. Для цього слід сполучити на гістограмі розподілу праву верхню вершину прямокутника, що відповідає модальному інтервалу (має найбільшу частоту) і з правою вершиною прямокутника, що відповідає передмодальному інтервалу (попередній прямокутник), а ліву вершину модального сполучити із лівою вершиною післямодального прямокутника (рис. 2.а).

Медіаною називається середнє значення ознаки в ранжованому ряді варіант, тобто значення рівновіддалене від початку і кінця варіаційного ряду. Якщо кількість варіант непарна, то медіаною є варіанта із порядковим номером:

$$M_e = \frac{x_{\frac{n+1}{2}}}{2} \tag{6}$$

Якщо кількість варіант парна, то:

$$M_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \tag{7}$$

Для того, щоб знайти медіану інтервального ряду, потрібно спочатку знайти ***медіанний інтервал***, тобто такий інтервал, для якого нагромаджена частота (або відносна частота) дорівнює півсумі всіх частот (відносних частот) або перевищує її. В загальному випадку значення медіани розраховується за формулою:

$$M_e = \underline{x}_{M_e} + h_{M_e} \cdot \frac{\frac{1}{2}\sum f - S_{M_e-1}}{f_{M_e}}, \tag{8}$$

де $\underline{x}_{M_e}$ – нижня межа медіанного інтервалу; h_{M_e} – довжина медіанного інтервалу; $\sum f$ – сума частот (відносних частот); S_{M_e-1} – сума частот, нагромаджених перед медіанним інтервалом; f_{M_e} – частота медіанного інтервалу.

Найлегше визначити медіану графічним способом. Для цього слід на кумуляті (кривій накопичених частот) поділити останню ординату (суму накопичених частот) навпіл і через точку з половинною максимальною ординатою(тобто накопиченою частотою – 50%) провести пряму, паралельно

осі Ох до перетину з кумулятою. Абсциса перетину цієї прямої із кумулятою і є медіаною (рис. 2.б).

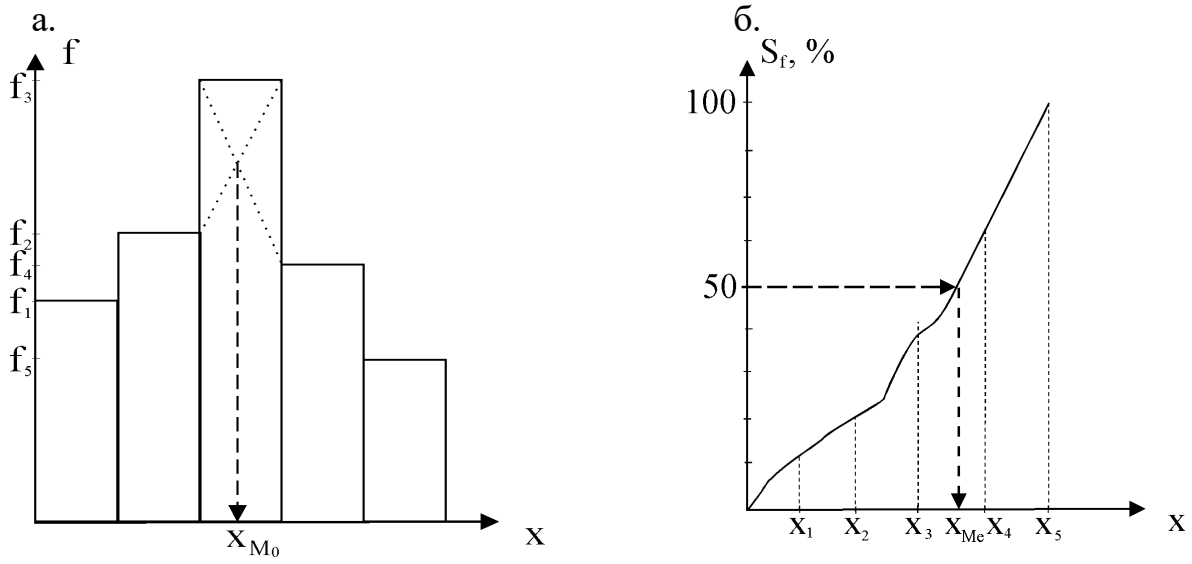


Рис. 2*. Графічний спосіб визначення моди (а) і медіани (б) інтервального ряду*

Характеристика розміру варіації. Середня величина є дуже інформативним показником. Але її типовість і надійність для всього ряду залежать від розміру і знаку відхилень пересічних варіант від середньої. Тому в статистиці важливим є не тільки розрахунок середньої, але й меж її довірчих інтервалів. Чим ці межі вужчі, тим краще середня характеризує загальний рівень ряду.

Відхилення значень варіант від середнього значення називається ***варіацією ряду***. Показники варіації поділяються на ***абсолютні*** і ***відносні***. Перші розраховуються за конкретними значеннями варіант. Серед абсолютних показників найважливішими є: розмах варіації, середнє арифметична відхилення і середнє квадратичне відхилення.

Розмах варіації являє собою максимальну амплітуду відхилення між максимальним і мінімальним значенням ряду:

$$R = x_{max} - x_{min}, \qquad (9)$$

де R – розмах варіації, x_{max} і x_{min} – відповідно найбільше і найменше значення ряду. Використовується розмах варіації, як правило, тоді, коли важливо знати інтервал можливого коливання значень ряду. Однак розмах варіації не дуже надійний показник, оскільки він залежить від значень крайніх у варіаційному ряду варіант, які, як правило, найменш надійні. Тому більш показовою є міра розсіяння варіант навколо середнього значення. Такій умові відповідають середнє арифметичне та квадратичне відхилення.

Середнє арифметичне відхилення розраховується за формулою:

$$\bar{d} = \frac{\sum |x - \bar{x}|}{n} \tag{10}$$

або те ж саме у зваженій формі:

$$\bar{d} = \frac{\sum |x - \bar{x}| \cdot f}{\sum f}, \tag{11}$$

де f – частота відповідної варіанти. Середнє арифметичне відхилення дає певне уявлення про варіацію ознаки, але має один суттєвий недолік: іноді буває так, що відхилення від середнього значення ряду є значними за модулем, а середнє відхилення не дуже велике. Тому суттєво більшого поширення набув показник ***середнього квадратичного відхилення***:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}, \tag{12}$$

або те ж саме у зваженій формі:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2 \cdot f}{\sum f}}, \tag{13}$$

де f – частота відповідної варіанти. Хоча і середнє арифметичне відхилення ($\bar{d}$) і середнє квадратичне відхилення (σ) характеризують один і той же ж параметр – варіацію, але між ними існує деяка розбіжність у числових значеннях. Це пояснюється тим, що при піднесенні до квадрата питома вага малих відхилень зменшується, а великих збільшується. В багатьох джерелах [фещур,] наводиться емпіричне відношення між цими величинами:

$$\sigma = 1{,}25\ \bar{d} \tag{14}$$

Відносних показників варіації відомо багато. В спеціальній літературі [фещур,] вони розглянуті дуже широко, але ми зупинемось лише на двох із них – ***коефіцієнтах варіації та дисперсії,*** що є найбільш вживаними. ***Коефіцієнт варіації*** являє собою відношення середнього квадратичного відхилення до середнього арифметичного варіаційного ряду, виражене в процентах, і розраховується за формулою:

$$V = \frac{\sigma}{\bar{x}} \cdot 100\% \tag{15}$$

Коефіцієнт варіації використовується для оцінки однорідності сукупності. Якщо $V \leq 33\%$, сукупність є однорідною, $\bar{x}$ – типовою і надійною характеристикою сукупності.

На основі вищесказаного ми підходимо до розуміння сутності ***дисперсії***.

Дисперсія – це середня арифметична квадратів відхилень варіант від їх середньої арифметичної:

$$\sigma^2 = \frac{\sum (x - \bar{x})^2 \cdot f}{\sum f} \tag{16}$$

Як видно з формули (13), середнє квадратичне відхилення, по суті, є коренем квадратним із дисперсії. Дисперсія володіє властивістю мінімальності: дисперсія менша середньої арифметичної квадратів відхилень варіант від любої сталої величини, що відрізняється від середньої арифметичної.

Якщо сукупність доволі велика і не дуже однорідна, то її, як правило, ділять на декілька груп. Тоді для оцінки впливу групоутворюючих ознак на загальну варіацію ознаки важливими є такі показники: загальна дисперсія, групова дисперсія, середня з групових дисперсій, міжгрупова дисперсія. ***Загальна дисперсія*** в такому випадку розраховуватиметься за формулою (16).

Групова дисперсія становитиме:

$$\sigma_i^2 = \frac{\sum (x - \bar{x}_i)^2 \cdot f}{\sum f} \tag{17}$$

середня з групових дисперсій:

$$\bar{\sigma}_i^2 = \frac{\sum \sigma_i^2 \cdot f_i}{\sum f_i} \tag{18}$$

а ***міжгрупова дисперсія***:

$$\delta_i^2 = \frac{\sum (\bar{x} - \bar{x}_i)^2 \cdot f_i}{\sum f_i} \tag{19}$$

де $\bar{x}_i$ – середня для *i*-тої групи; $\bar{x}$ – середня для всієї сукупності; f – частота реалізація явища; f_i – обсяг *i*-тої групи.

Якщо групові дисперсії не дорівнюють нулю, то значить в межах виділених груп ще залишилась варіація, зумовлена зовнішніми по відношенню до групувальних ознак. Така варіація оцінюється ***залишковою дисперсією***. Міру впливу групувальної ознаки на утворення загальної дисперсії характеризує міжгрупова дисперсія. Тому ***загальна дисперсія*** дорівнює сумі середньої з внутрішньогрупових дисперсій і міжгрупової дисперсії:

$$\sigma^2 = \bar{\sigma}^2 + \delta^2 \tag{20}$$

Відношення міжгрупової дисперсії до загальної дисперсії називають ***коефіцієнтом детермінації***:

$$\eta^2 = \frac{\delta^2}{\sigma^2} \qquad (21)$$

Цей коефіцієнт показує яку частину загальної дисперсії становить міжгрупова дисперсія, або іншимии словами наскільки дія групувальної ознаки (тобто покладеної в основу виділення груп) впливає на загальну варіацію досліджуваної ознаки. Коефіцієнт детермінації використовується для оцінки адекватності рівняння тренду при побудові регресійних моделей екологічних явищ та процесів. Це дуже зручно використовувати при роботі з табличним процесором Excel 2000 пакету прикладних програм Microsoft Office 2000 (або іншими версіями цієї програми). В даній комп'ютерній програмі коефіцієнт детермінації позначається R^2. Якщо програма пропонує декілька альтернативних ліній тренду, то адекватнішим буде той тип рівняння тренду, для якого величина буде R^2 більшою.

а. б.

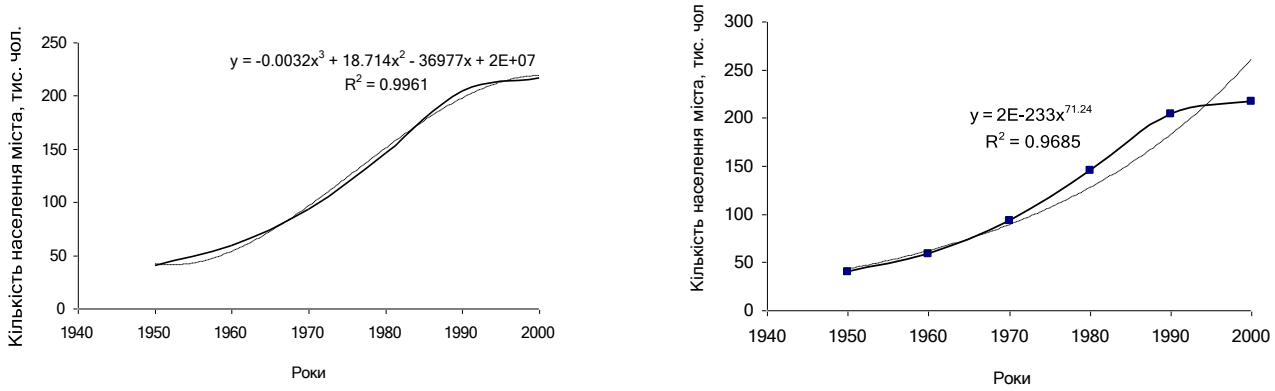


Рис. 3. *Побудова лінії тренду за допомогою різних функцій (поліноміальної (а) та степеневої (б)) за допомогою табличного процесора Excel 2000*

Як видно із рис. 3 більш адекватно описує рівняння лінії тренду графік поліноміальної функцій, тому для нього характерно вище значення коефіцієнта детермінації.

Ми навмисно так широко зупинились на розрахунових формулах різних видів дисперсії. Дисперсійний аналіз є одним із найбільш поширених у моделюванні та прогнозуванні стану довкілля методів статистики. Він дозволяє відповісти на питання: чи вірогідний вплив того чи іншого фактора на стан техноекосистеми, або на результати впровадження тих чи інших природоохоронних заходів. Він також дає можливість порівнювати між собою декілька системно зв'язаних вибірок і визначити, чи існують між ними статистично вірогідні відмінності та яка ймовірність цих відмінностей. Спільною рисою всіх дисперсійних моделей є те, що у них перевіряється дія деякого загального фактора (в багатофакторному аналіз – дія одночасно кількох

взаємопов'язаних факторів) на об'єкт (техноекосистему). Детальніше про дисперсійний аналіз можна прочитати у [,].

І накінець, третьою важливою групою показників аналізу варіаційного ряду є ***характеристики форми розподілу***. Якщо безкінечно зменшувати інтервал варіаційного ряду, то з часом гістограма розподілу перетвориться на плавну криву, яка характеризуватиме залежність частоти реалізації події від величини досліджуваної ознаки. Таку криву називають ***функцією щільності розподілу***, або просто ***щільністю***, або ***розподілом***. Основними характеристиками форми розподілу є число вершин розподілу, асиметрія та ексцес.

Число вершин. Вершина кривої розподілу, по суті, є модою. Розподіл з однією вершиною називається ***унімодальним*** (мономодальним), з двума – ***бімодальним***, з багатьма вершинами – ***мультимодальним***.

Асиметрія. Деякі криві розподілу (наприклад, при нормальному розподілі) симетричні відносно вертикальної лінії, що проходить через моду. Іноді одна сторона кривої більш полога ніж інша. В такому випадку кажуть про ***асиметрію*** (скошеність) розподілу. Асиметрія визначається за формулою:

$$A_s = \frac{\sum(\overline{x_i} - \bar{x})^3 \cdot f}{\sigma^3 \sum f} \qquad (22)$$

При нормальному розподілі $A_s = 0$, при правосторонній симетрії $x > \bar{x}$ для більшості варіант, тому $A_s > 0$, а при лівосторонній симетрії навпаки – $A_s < 0$ (рис. 4). Іноді правосторонню симетрію називають ще ***додатньою***, а лівосторонню – ***від'ємною***.

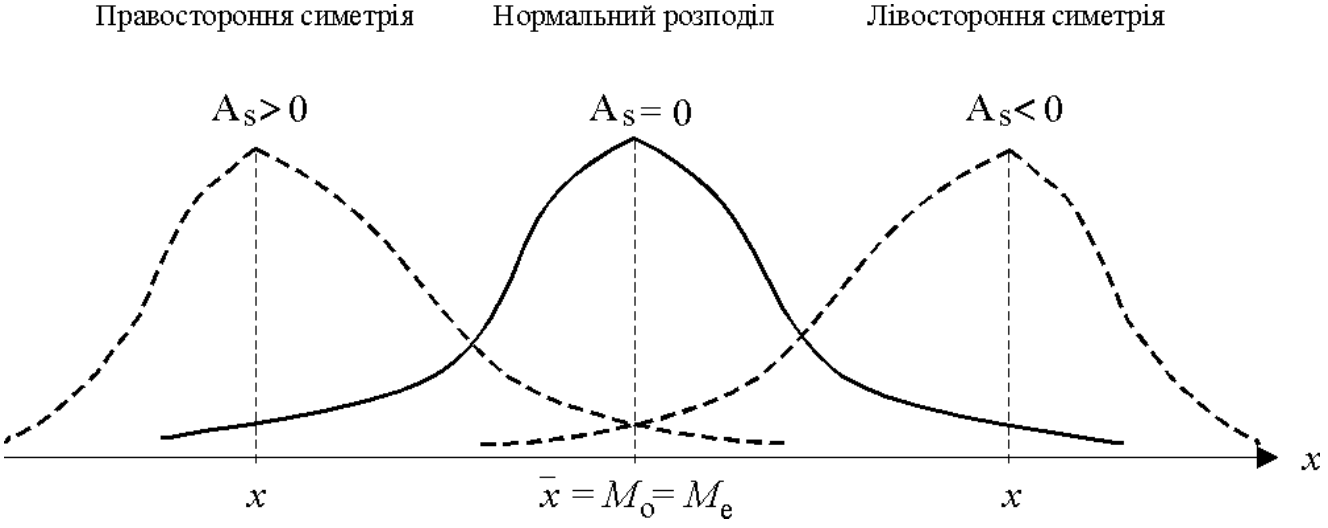


Рис. 4. *Асиметрія розподілу*

Якщо розподіл симетричний, то для нього можна розрахувати показник ***ексцесу***:

$$E_x = \frac{\sum (x - \bar{x})^4 \cdot f}{\sigma^4 \sum f} \qquad (23)$$

Для вищерозглянутого випадку нормального розподілу $E_x = 0$, при $E_x > 0$ маємо ***гостровершинний*** розподіл, при $E_x < 0$ – ***плосковершинний*** розподіл. Додатнє значення E_x вказує на те, що крива щільності має в околі моди (медіани) більш високу і гострішу вершину. В цьому випадку кажуть про ***додатній*** ексцес у порівнянні із нормальною кривою. Від'ємне значення E_x (***від'ємний*** ексцес) зумовлює нижчий і пологіший характер вершини у порівнянні із нормальним розподілом (рис. 5).

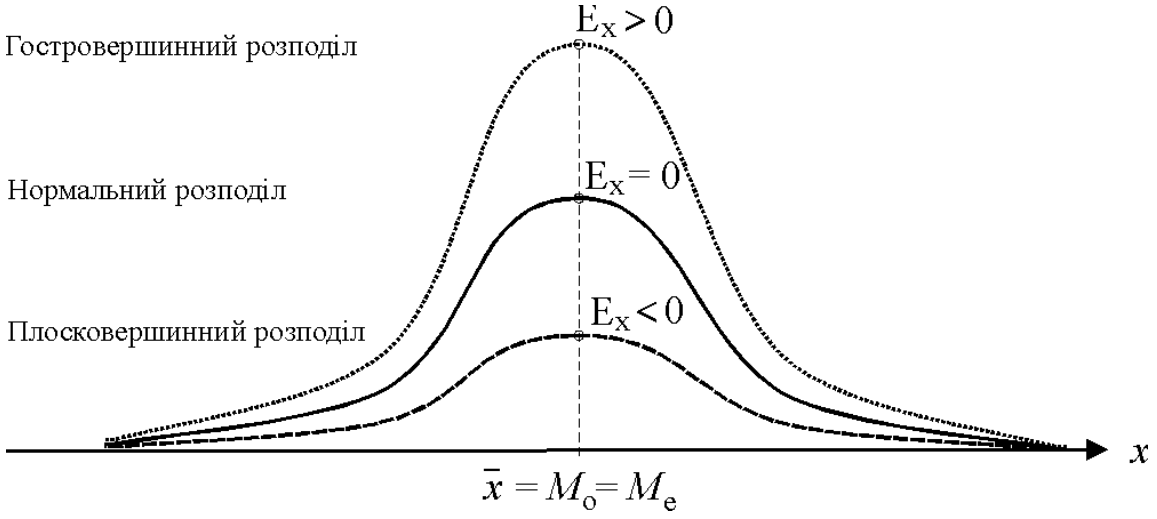


Рис. 5. *Ексцес розподілу*

Проаналізувавши таким чином варіаційний ряд, тобто визначивши його характеристики центру розподілу, розміру варіації та форми розподілу, можна приступити до моделювання ряду розподілу. ***Моделювання ряду розподілу*** теж здійснюється у три етапи:

- дослідження загальних характеристик розподілу;
- вирівнювання (підгонка) емпіричного розподілу за теоретичною кривою розподілу;
- перевірка узгодженості теоретичного розподілу емпіричному.

Дослідження загальних характеристик розподілу включає в себе аналіз варіаційного ряду. Іншими словами початковий етап побудови моделі розподілу полягає у визначенні середніх значень варіант варіаційного ряду та окремих інтервалів інтервального ряду, моди, медіани, середнього квадратичного відхилення, дисперсії, асиметрії та ексцесу розподілу. Після цього здійснюється побудова графічного зображення розподілу у вигляді одного із типів графіків (рис. 1).

На наступному етапі моделювання вибирається вид теоретичного розподілу, а потім здійснюється вирівнювання емпіричного ряду до вибраного виду теоретичного розподілу. Теоретичний розподіл відображає лише загальні

закономірності розподілу, що проявляються у вигляді певних внаслідок дії лише основних факторів. Тобто теоретичний розподіл описує в генералізваному аналітичному вигляді функціональний зв'язок між значеннями варіант ряду та їх частотами. В математичній статистиці та теорії ймовірності відомо багато стандартних видів розподілу: нормальний розподіл, біноміальний розподіл, розподіл Пуассона, Бета-розподіл, розподіл Коші, розподіл хі-квадрат, логнормальний розподіл, розподіл Стьюдента тощо. В спеціальній літературі [, ,] можна взнати детальніше про дані розподіли. Ми зупинемось на розгляді лише перших трьох розподілів, оскільки саме вони є найбільш вживаними.

Нормальний розподіл визначає щільність розподілу безперервної випадкової величини для значення варіант:

$$P(m) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}, \qquad (24)$$

де P (m) – щільність розподілу (іноді це ще називають диференціальною функцією Лапласа), t – нормоване відхилення ($t = \frac{x - \bar{x}}{\sigma}$), σ – стандартне (середнє квадратичне відхилення).

Він можливий, коли на значення варіант впливає велика кількість незалежних випадкових факторів і жоден із них не є жорстко домінуючим для даної ознаки. Нормальний розподіл описується симетричною кривою (рис. 4) і володіє рядом властивостей, що роблять його практично найбільш вживаним у статистиці. Так зокрема, площа, обмежена кривою щільності розподілу, дає загальну ймовірність реалізації події, що дорівнює 1, а площа під кривою між двума фіксованими точками пропорційна ймовірності появи точок, що знаходяться в проміжку між ними. Звідси, чим менше стандартне відхилення, тим гостріший пік кривої розподілу і тим більша кількість точок розміщуватиметься поблизу середнього значення.

Це дозволяє оцінити ймовірність даної події: 68,26% площі під кривою потрапляє в інтервал від $\bar{x} - \sigma$ до $\bar{x} + \sigma$, 95,44% площі зосереджено між $\bar{x} - 2\sigma$ і $\bar{x} + 2\sigma$, а 99,73 – між $\bar{x} - 3\sigma$ і $\bar{x} + 3\sigma$. Крива нормального розподілу асимптотично наближається до осі *x* і для того, щоб заповнити весь простір під нею потрібна величезна кількість емпіричних спотережень. Але, як випливає із вищесказаного, за межами інтервалу від $\bar{x} - 4\sigma$ до $\bar{x} + 4\sigma$ лежить настільки мала частина площі, що ймовірностями, які не вкладаються в даний інтервал на практиці найчастіше нехтують.

Біноміальний розподіл поряд із нормальним теж належить до найбільш простих і застосовуваних розподілів. За біноміальним законом ймовірність настання певної події *m* разів при *n* незалежних спробах визначається за формулою:

$$P_m^n = \frac{n!}{m!(n-m)!} \cdot p^m \cdot (1-p)^{n-m}, \qquad (25)$$

де p – ймовірність реалізації певної події;

m – частота її реалізації;

n – загальна кількість подій.

Біноміальний розподіл в моделюванні і прогнозуванні стану довкілля найчастіше застосовується, коли потрібно виявити закономірності розподілу ймовірності прояву двох взаємовиключаючих екологічних явищ при заданому рівні ймовірності *p*. Типовий приклад: нехай в басейні річки знаходиться певне промислове підприємство. За рік на ньому виникає 6 аварійних ситуацій із скидом забруднених стічних вод у річку. Усі ці скиди забруднених вод різні за розмірами і екологічною шкодою. В результаті деяких із таких скидів відбувається розчинення забруднюючих речовин у річковій воді і забруднення річки цими речовинами до концентрації, що перевищує ГДК. В цьому випадку позначимо кількість аварійних ситуацій через *n*, а кількість таких скидів, що призвели до забруднення води річки до рівня, вище ГДК – *m*. В таблиці 3 наведено результати розрахунку ймовірностей за біноміальним законом при *p* = 0,5; *n* = 6. Як видно з рис. 6.. графік біноміального розподілу є дискретним.

Таблиця 3. *Біноміальний розподіл (p = 0,5; n = 6).*

m	P_m^n
1	0,097
2	0,237
3	0,314
4	0,237
5	0,097
6	0,019

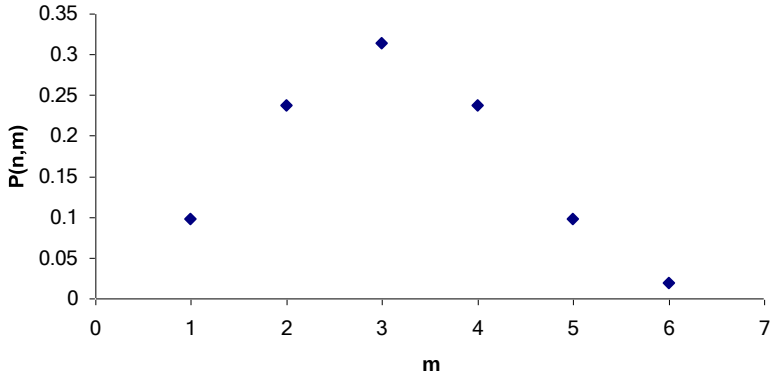


Рис. 6. *Графік біноміального розподілу*

Розподіл Пуассона являє собою по суті граничний випадок біноміального розподілу, коли ймовірність прояву явища дуже мала, а загальний розмір вибірки – великий. Ним користуються, зазвичай, при описі частот розподілу дуже рідких подій, наприклад, катастрофічних повеней чи паводків за довгого періоду спостережень. Щільність розподілу задається співвідношенням:

$$P_m = \frac{\mu^m \cdot e^{-\mu}}{m!}, \tag{26}$$

де μ – середнє число подій, яке визначається як добуток кількості варіант на ймовірність їх прояву в одноразовій спробі ($\mu = n \cdot p$); m – частота прояву події.

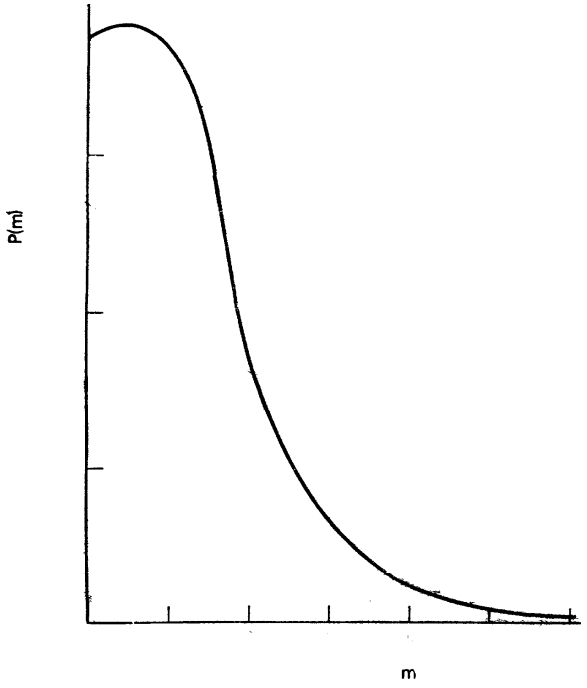


Рис. 7. *Щільність розподілу Пуассона*

Як видно з графіка на рис. 7, розподіл Пуассона дуже близький до асиметричного біноміального.

Він застосовується при аналізі рідких і незалежних одне від одного явищ в часі і просторі.

Так, наприклад, даний вид розподілу використовується при побудові моделі частоти прояву катастрофічних паводків. До Пуассонівського наближається також і розподіл мікроелементів в зразку грунту. Аналогічний розподіл успішно застосовується при гідрохімічному аналізі природних вод, коли мова не йде про детерміновані моделі розсіювання забруднень, оскільки при їх побудові використовуються зовсім інші методи.

Наступним етепом є ***вирівнювання (підгонка) емпіричного розподілу за теоретичною кривою розподілу***. Вона зводиться до порівняння порівняння частот фактичного розподілу з відповідними теоретичними частотами. Теоретичні частоти розраховуються за формулами (24-26). Наприклад, для нормального розподілу вони розраховуються за співвідношенням:

$$f_{теор.} = P(m) \cdot \frac{n \cdot h}{\sigma}, \tag{27}$$

де *P(m)* – значення диференційної функції Лапласа (розраховується за формулою (24)); *n* – загальне число спостережень, *h* – ширина інтервалу; *σ* – стандартне (середнє квадратичне) відхилення.

Після цього приступають до наступного етапу моделювання рядів

розподілу – ***перевірки узгодженості теоретичного розподілу емпіричному***. Вона здійснюється за допомогою ***критеріїв узгодженості.*** Їх існує доволі багато – критерій Пірсона (χ^2, типу I, III, IV, VI), Романовського, Колмогорова, Ястремського, так званий критерій "$n\omega^2$" і т.д. У зв'язку із напрямком і задачами нашої роботи зупинимось лише на критеріях Пірсона χ^2 і Колмогорова.

Англійський вчений ***К. Пірсон*** запропонував ***критерій***, статистичну характеристику якого розраховують за формулою:

$$\chi^2 = \sum \frac{(f - f')}{f'}, \tag{28}$$

де f і f' – відповідно фактичні і теоретичні частоти.

За спеціальними таблицями [] визначають ймовірність досліджуваного значення χ^2 залежно від числа ступенів свободи (вільності) – $k = m - 3$, де k – число ступенів свободи, а m – кількість інтервалів (груп) і заданому рівні значимості, яка в західній спеціальній літературі та комп'ютерних програмах позначається *p-level*. Рівень значимості приймається як правило 0,05 або 0,01.

Якщо розраховане за формулою (28) значення критерію менше за табличне, то при прийнятому рівні значимості відхилення між фактичними і розрахунковими частотами несуттєве і спричинене випадковими факторами, а отже емпіричний розподіл відповідає теоретичному.

Критерій Колмогорова розраховується за наступною формулою:

$$\lambda = \frac{\max\left| S_f^{теор.} - S_f^{емп.} \right|}{\sqrt{n}} \tag{29}$$

λ – критерій Колмогорова, $S_f^{теор.}$ і $S_f^{емп.}$ – сума накопичених теоретичних і емпіричних частот в долях одиниці, n – сума емпіричних частот.

Значення даного критерію теж табульовані [ ]. За заданого рівня значимості $\alpha = 1 - k\ (\lambda)$ по цих таблицях визначаються критичні значення λ_α Якщо розраховане значення критерію λ менше критичного λ_α, то приймається гіпотеза про відповідність емпіричного розподілу теоретичному.

Критерій Колмогорова використовується для перевірки відповідності при достатньо великому n, але сучасні дослідження довели, що вже при $n \geq 20$, розподіл статистики вже практично не залежить від n.

Даний критерій узгодженості визначається фактично по одній точці, а тому в деяких випадках може не відображати загальну узгодженість теоретичного та емпіричного розподілу. Крім того найбільша різниця накопичених теоретичних та емпіричних частот, як правило, спостерігається в середній частині кривої розподілу. Тому судження про співпадання теоретичних та емпіричних розподілів в крайових частинах (а іноді саме це потрібно в практичних розрахунках) потрібно виносити дуже обережно.

2. Аналіз структури та дослідження взаємозв'язків у екологічних системах

В моделюванні і прогнозуванні стану довкілля велика увага приділяється аналізу взаємозв'язків між різноманітними природними та техногенними процесами та явищами. Більшість екологічних процесів формується ланцюгом причинно-наслідкових явищ, що змінюються в часі і просторі. Для вивчення цих процесів необхідно встановити їх причини, рушійні сили або джерела, тенденції розвитку. При цьому виникає необхідність в дослідженні залежностей, що пов'язують досліджувані процеси між собою та з іншими процесами і явищами. Наприклад, для моделювання забруднення річкових вод важливим питанням є коливання стоку річки (витрати води). При розрахунках і прогнозах стоку слід дослідити і з'ясувати зв'язки між коливаннями водності річки та коливанням шару опадів (дощових і снігових), температури повітря, випаровуваності з поверхні суші і водойм, запасів води в сніговому покриві і т.д.

Задачі по дослідженню структури та взаємозв'язків екологічних процесів дуже різноманітні. Умовно їх можна розділити на дві групи:

- задачі, пов'язані із виясненням причин або факторів розвитку процесів в часі і просторі;
- задачі, пов'язані із визначенням конкретних значень або тенденцій розвитку даного процесу в майбутньому.

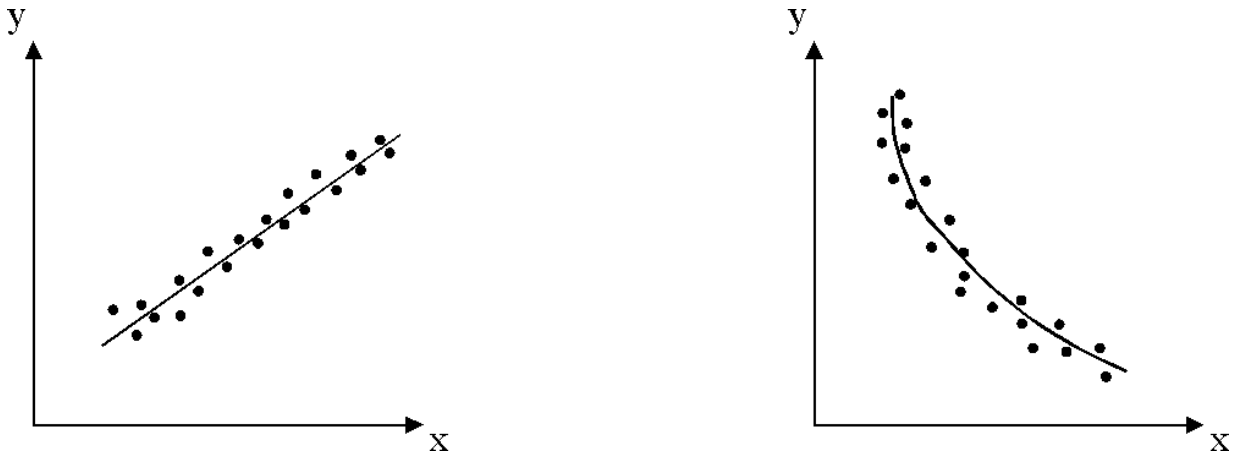


Рис. 8. Графіки прямолінійного та криволінійного зв'язку

Слід зауважити також, що задачі дослідження залежностей при моделюванні, взагалі то кажучи, не є чисто математичними задачами. Оскільки математичний аналіз в даному випадку обов'язково повинен супроводжуватись фізичним аналізом сутності екологічних процесів, відсутність останнього ж може спричинити серйозні прорахунки, похибки, а іноді й просто парадоксальні результати.

Взаємозв'язок досліджуваних процесів оцінюється по відповідності змін їх значень в часі і просторі. За формою графіка ця залежність може бути ***прямолінійною*** або ***криволінійною*** (рис. 8). Прямолінійні залежності вивчені краще, але в природі зустрічаються рідше. За тіснотою зв'язку або ступенем

визначеності одного із співставлюваних процесів відносно іншого, зв’язки можуть бути ***функціональні*** і ***стохастичні***. Окремим випадком даної класифікації є ***відсутність зв’язку*** між досліджуваними процесами.

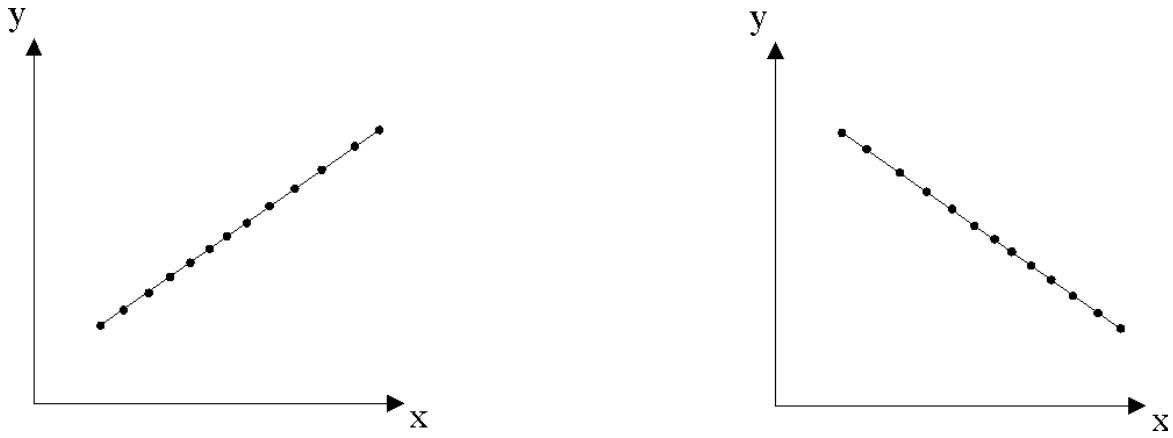


Рис. 9. Графіки зв’язку функціонально залежних величин

Функціональні зв’язки. Іноді залежність між досліджуваними величинами буває настільки тісною, що знаючи значення однієї із величин можна вказати точні значення іншої. Такі взаємозв'язки називаються функціональними. Наприклад, при моделюванні процесів забруднення повітря та масопереносу в атмосфері, для розуміння фізичної суті процесу, потрібно знати основні закони термодинаміки. Один із цих законів пов’язує тиск (*P*), температуру (*T*) і об’єм (*V*) газу:

$$P = \frac{R \cdot T}{V} \tag{30}$$

За цим законом кожному індивідуальному значенню *V* при фіксованих значеннях *R* і *T* відповідає лише одне значення *P*. Якщо нанести відповідні пари значень співставлюваних величин на графік в декартових координатах (рис. 9) у вигляді точок, то утворена група точок лежатиме строго на одній лінії, тобто кожному значенню *V* відповідатиме певне значення *P* і навпаки.

Такі залежності зустрічаються, як правило, в технічних науках, у фізиці, екологічні явища виявляються пов’язані функціональною залежністю доволі рідко.

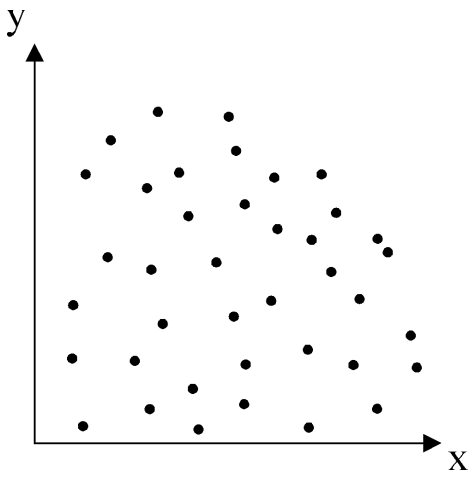


Рис. 10. *Графік зв'язку незалежних величин*

Відсутність зв'язку. Досліджувані екологічні процеси і явища іноді бувають непов'язані одне з одним, тобто зміна одного із них відбувається незалежно від зміни іншого.

Випадкова величина Y називається незалежною від випадкової величини X, якщо закон розподілу Y не залежить від того, яке значення значення набуває X.

Графік зв'язку незалежних величин (рис. 10) являє собою поле точок, у якому кожному значенню X відповідає з відповідною щільністю весь діапазон можливих значень Y і навпаки.

Ймовірнісна (стохастична) залежність. Зустрічається в екології найчастіше. Якщо Y пов'язане із X стохастичною залежністю, то знаючи значення X не можна вказати точне значення Y, а можна вказати лише закон розподілу величини Y, який залежить від величини X.

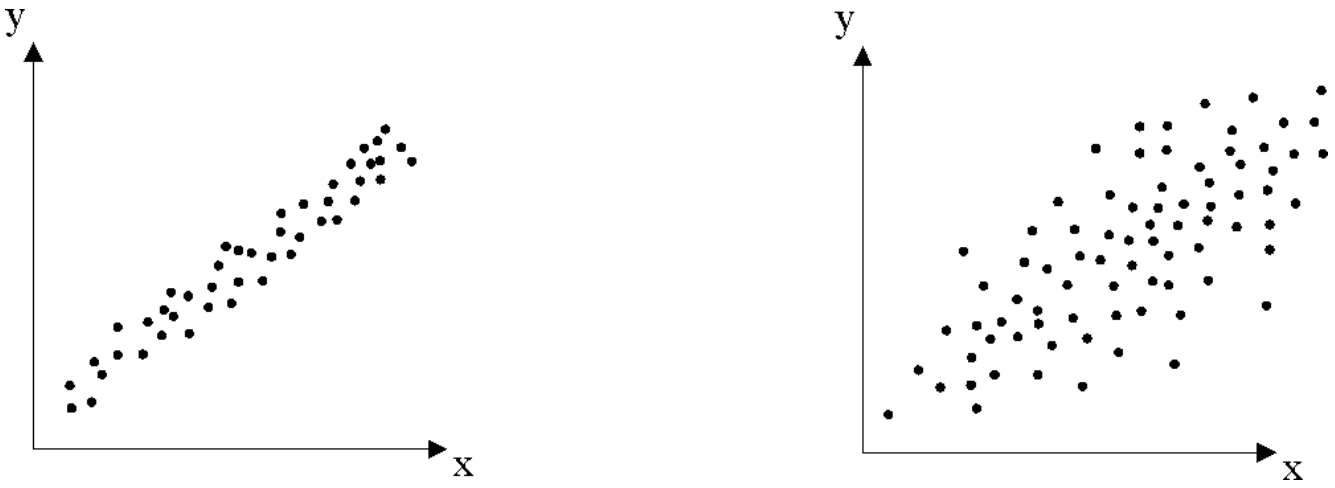


Рис. 11. *Графіки зв'язку стохастично залежних величин із високою та низькою щільністю зв'язку.*

На рис. 11 зображено графік стохастичної залежності більшої та меншої щільності зв'язку. Як видно із рисунку, точки кожної пари значень (X;Y) розміщуються в полі графіка ніби відносно якоїсь лінії. При цьому кожному значенню X відповідає своя смуга значень Y. Разом із зміною X змінюється середнє із можливих значень Y при даному X.

Стохастичний зв'язок між змінними проявляється, як правило, тоді, коли поряд із факторами, які впливають або на один або на інший процес, існують фактори, що впливають спільно як на перший, так і на другий процес.

По мірі збільшення щільності зв'язку форма графіка наближатиметься до графіка функціональної залежності. Таким чином, функціональна залежність є

крайнім випадком найбільш тісної стохастичної (ймовірнісної) залежності. Полярним випадком є повна незалежність змінних. Між цими граничними випадками розміщуються всі градації ймовірнісної залежності – від найбільш сильної до найслабшої.

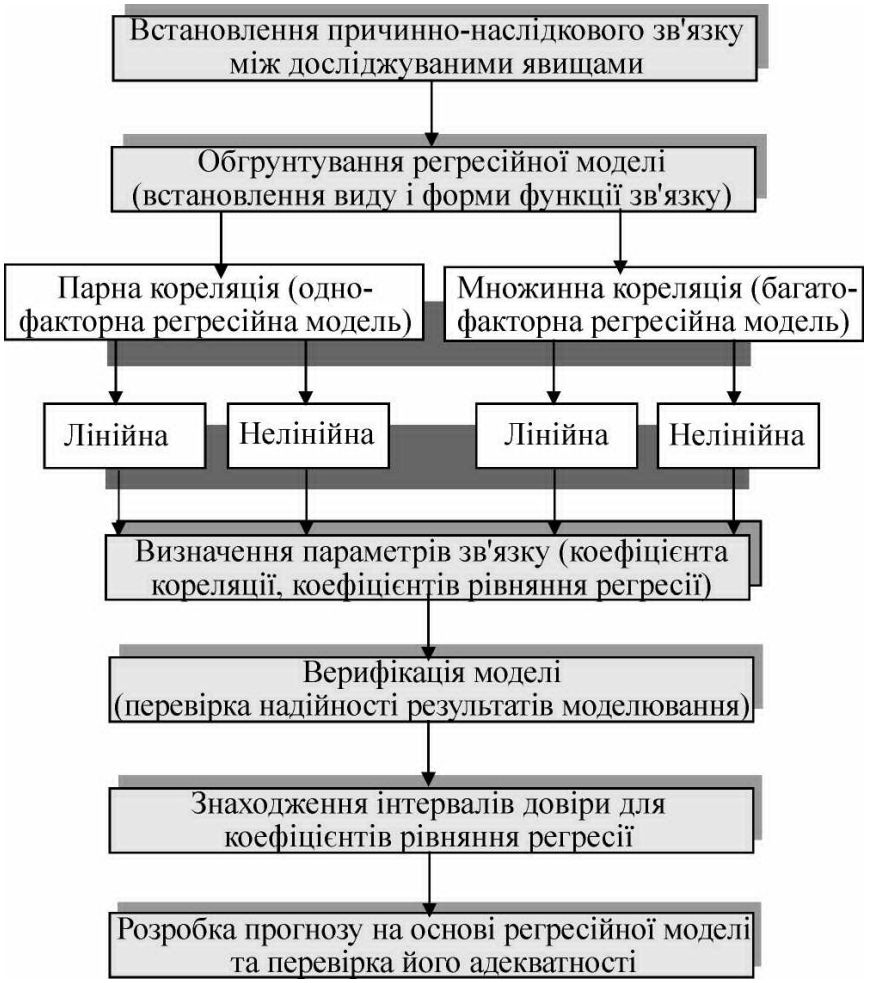


Рис. 12. *Алгоритм реалізації моделі кореляційно-регресійного зв'язку*

Регресійна модель дослідження взаємозв'язку між екологічними процесами. Для дослідження взаємозв'язку двох екологічних процесів або явищ найчастіше застосовується математична модель у вигляді рівняння регресії. Така модель називається ***регресійною*** або ***кореляційно-регресійною***. Процес її побудови в загальному складається з двох етапів:

- встановлення на основі великої кількості спостережень того, як змінюється в середньому функція Y в залежності від зміни одного або декількох її головних агрументів (іншими словами – визначення форми зв'язку і знаходження рівняння зв'язку двох змінних величин);
- визначення ступеня взаємозв'язку двох досліджуваних явищ (якщо ці явища взаємопоєднані) або ступеню впливу головних досліджуваних факторів на досліджуваний вплив (якщо ці зв'язки носять причинно-наслідковий характер).

В роботі [ ] наводиться більш розгорнутий алгоритм реалізації моделі кореляційно-регресійного аналізу (рис. 12).

Однофакторна лінійна регресійна модель. Зупинимось детальніше на першому етапі дослідження взаємозв'язку двох змінних величин – ***встановленні рівняння взаємозв'язку в лінійному вигляді***. Як відомо з попередніх розділів, загальний вигляд лінійного зв'язку:

$$\bar{y} = ax + b, \tag{31}$$

де $\bar{y}$ – середнє із можливих значень *Y* при даному *x*. Функція, що виражає зв'язок між значенням аргументу і умовним середнім арифметичним досліджуваної залежної змінної, називається ***рівнянням лінії регресії***. До даного рівняння, поряд із змінними, входять і коефіцієнти – *a* і *b*. Зміст цих коефіцієнтів детально розглядався в розділі, присвяченому використанню елементарних функцій в моделюванні і прогнозуванні стану довкілля. Так, зокрема, коефіцієнт *a* називається ще ***кутовим коефіцієнтом*** і характеризує нахил прямої лінії (графіка функції) до осі абсцис. Математичний зміст цього коефіцієнта полягає у тому, що він характеризує нахил лінії графіка функції до осі абсцис і дорівнює: $k = tg\ \alpha$. Коефіцієнт *b* називають ***вільним членом рівняння***. Він показує довжину відрізка, який відсікає лінія графіку від початку координат. Якщо рівняння функції не містить коефіцієнта *b*, то її графік пройде через початок координат.

Для знаходження параметрів парної регресії затосовують ***метод найменших квадратів*** (МНК). Суть методу полягає в тому, що сума квадратів відхилень x_i від їх середньої $\bar{x}$ є величиною мінімальною:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min \tag{32}$$

За цим методом розрахований теоретичний розподіл має максимально точно відповідати емпіричному, тобто точки лінії регресії, розраховані теоретичним шляхом, повинні бути отримані таким чином, щоб сума їх відхилень від емпіричних (дослідним шляхом добутих) значень була мінімальною:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 = \min \tag{33}$$

Знайдемо параметри a і b для (31) за допомогою методу найменших квадратів. Позначимо різницю фактичних і розрахованих значень через *S*:

$$S = y_i - \bar{y}(x_i) = y_i - ax_i - b \tag{34}$$

Тоді:

$$S^2 = (y_i - \bar{y}(x_i))^2 = (y_i - ax_i - b)^2 \to \min \tag{35}$$

Для знаходження мінімуму функції S^2 від параметрів a і b слід визначити частинні похідні по a і b. Вони відповідно дорівнюють:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2(\sum xy - a\sum x^2 - b\sum x \\ \dfrac{\partial S}{\partial b} = -2(\sum y - a\sum x - bn \end{cases} \qquad (36)$$

Як відомо, функція досягає мінімуму якщо її перша похідна дорівнює 0. Прирівнявши частинні похідні (36) до нуля, отримаємо систему двох лінійних рівнянь із двума невідомими (a і b). Цю систему іноді ще називають ***нормальною системою рівнянь*** або ***системою Гаусса***:

$$\begin{cases} a\sum x^2 + b\sum x = \sum xy \\ a\sum x + bn = \sum y \end{cases} \qquad (37)$$

Розв'язавши цю систему отримаємо:

$$\begin{cases} a = \dfrac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2} \\ b = \dfrac{\sum y - a\sum x}{n} = y - ax \end{cases} \qquad (38)$$

Коефіцієнти лінійної регресії розраховуються за ретроспективний період спостережень для вибіркової сукупності. Тому для доведення репрезентативності вибірки слід виконати оцінку значимості коефіцієнтів регресії, тобто довести, що лінійний зв'язок у вибірковій сукупності свідчить про такий же ж зв'язок у генеральній сукупності. Значимість коефіцієнта лінійної регресії перевіряють за допомогою t-критерію Стьюдента. За таблицею t-розподілу Стьюдента [ ] знаходять величину $t_{\frac{\alpha}{2}}$ з $k = n - 2$ ступенями свободи і рівнем значимості α.

Розрахункове значення критерію знаходять за формулою:

$$t_{розр.} = |a| \cdot \sqrt{\frac{\sum (x - \bar{x})^2}{\sum (y - \bar{y})^2} \cdot (n - 2)} \qquad (39)$$

Якщо $t_{розр.} > t_{крит.}$, то коефіцієнт регресії вважається значимим, тобто характер зв'язку у вибірковій сукупності відповідає генеральній.

Але випадки лінійної регресії зустрічаються в моделюванні і прогнозуванні стану довкілля доволі рідко. Частіше лінійна регресія використовується для лінійної апроксимації існуючих екологічних закономірностей і взаємозалежностей.

Однофакторна нелінійна регресійна модель. В окремих випадках

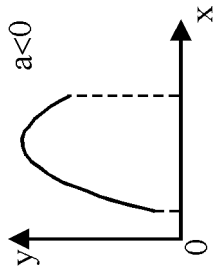
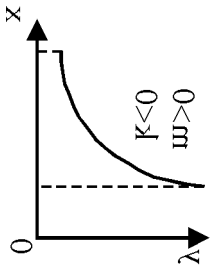
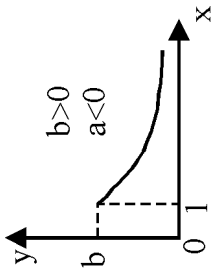
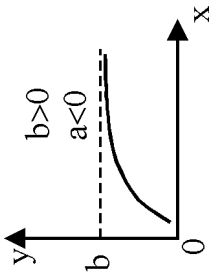
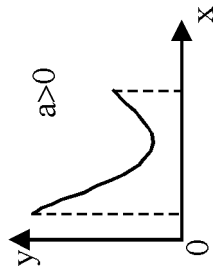
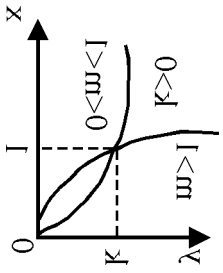
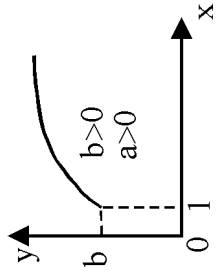
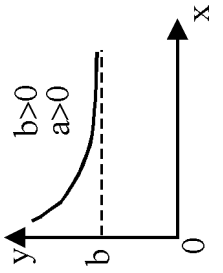
використати таку апроксимацію неможливо або ж недоцільно. Тоді досліджується ***модель парної нелінійної регресії***. Методика розрахунку коефіцієнтів парної нелінійної регресії аналогічна розгялянутій вище. В спеціальній літературі [,] вона описана доволі грунтовно. У зв'язку із напрямом та специфікою нашої роботи ми опустимо їх виведення і наведемо лише кінцеві формули для розрахунку коефіцієнтів парної нелінійної регресії у вигляді таблиці 4.

На наступному етапі на основі розрахунку параметрів рівняння регресії (лінійної чи нелінійної) вирішуються задачі, пов'язані із ***визначенням конкретних значень або тенденцій розвитку даного процесу в майбутньому***. ***Прогнозуванням*** називається наукове передбачення ймовірнісних шляхів розвитку явищ і процесів для більш-менш віддаленого майбутнього. Проміжок часу від моменту, для якого є останні статистичні дані про досліджуваний об'єкт, до моменту, до якого належить прогноз, називається ***періодом упередження***.

Суть прогнозування на основі регресійної моделі полягає у ектраполяції на майбутнє розрахованих за попередні періоди залежностей. Такий метод прогнозування виходить із збереження загальної тенденції розвитку явищ (процесів) у часі. На практиці прогноз показника отримують підстановкою у здобуте рівняння конкретного значення детермінуючого (визначаючого) фактора. Результатом прогнозу є точкова оцінка середнього значення функції при заданому рівні прояву (реалізації) фактора.

Таблиця 4. *Розрахунок коефіцієнтів парної нелінійної регресії (наводиться за [] із змінами авторів)*

Система рівнянь для визначення коефіцієнтів	$a\sum x^2+b\sum x+cn=\sum y$ $a\sum x^3+b\sum x^2+c\sum x=\sum xy$ $a\sum x^4+b\sum x^3+c\sum x^2=\sum yx^2$	$n\lg k+m\sum\lg x=\sum\lg y$ $\lg k\sum\lg x+m\sum\lg^2 x=\sum\lg x\sum\lg y$	$a\sum\lg x+bn=\sum y$ $a\sum\lg^2 x+b\sum\lg x=\sum y\lg x$	$a\sum\frac{1}{x}+bn=\sum y$ $a\sum\frac{1}{x^2}+b\sum\frac{1}{x}=\sum\frac{y}{x}$

Графік рівняння				
				
				
Рівняння регресії	$y = ax^2 + bx + c$	$y = kx^m$	$y = a \cdot \lg x + b$	$y = a/x + b$

Середнє значення прогнозу показника (вислідної ознаки) $\bar{y}_{np.}$ при значенні детермінючого фактора $\bar{x}_{np.}$, відповідно до рівняння лінійної регресії, визначається за формулою:

$$\bar{y}_{np.} = a\bar{x}_{np.} + b \tag{40}$$

Окремим важливим питанням є достовірність прогнозу. Для того аби довести достовірність прогнозу слід визначити межі його довірчого інтервалу $\Delta\bar{y}_{np.}$:

$$\Delta\bar{y}_{np.} = t_{a,k} \cdot S \cdot \left[(1 + \frac{1}{n} + \frac{(x_{np.} - x)^2}{\sum_{i=1}^{n}(x_i - x)^2}) \right]^{\frac{1}{2}} \tag{41}$$

де $t_{a,k}$ – *t*-розподіл Стьюдента з $k = n - 2$ ступенями свободи і рівнем значимості α; n – кількість варіант (спостережень), S – точкова оцінка величини дисперсії σ.

Багатофакторна регресійна модель. Природа являє собою динамічну,

функціональну і поліваріантну систему, що об'єднана в єдине ціле надзвичайно складним сплетінням різноманітних постійно взаємодіючих природних факторів. Доволі часто в природі зустрічаються випадки, коли значення параметрів техноекосистеми формуються під впливом не однієї а кількох різних факторних ознак. При цьому жодна із даних ознак не справляє вирішального впливу на досліджувані параметри системи, але їх спільний вплив є відчутним. В такому випадку для дослідження структури взаємозв'язків у ТЕС використовують модель багатофакторної регресії.

Найскладнішим моментом багатофакторного регресійного моделювання є те, що, по-суті, досліджується зв'язок між значенням залежної змінної та довільною комбінацією незележних (факторних) змінних. У зв'язку із такою поліваріантністю, складністю побудови моделі багатофакторної регресії та вибору кінцевої функції найчастіше для багатофакторного регресійного аналізу використовуються ***метод виключень***, ***покроковий регресійний аналіз*** і ***метод всіх можливих регресій*** [ ].

Суть ***методу виключення*** полягає в тому, що будують рівняння, яке містить всі фактори. Потім для кожного фактора обчислюють статистику часткового *F*-критерію, тобто відношення різниці суми квадратів всіх факторів і всіх крім останнього фактора до дисперсії часткової величини цієї моделі. Якщо мінімальне значення часткового *F*-критерію менше за критичне (взяте із таблиці), то даний фактор виключається із моделі. Процедура продовжується поки не будуть виключені всі фактори, що відповідають даній умові.

Покроковий регресійний аналіз являє собою зворотну процедуру. Спочатку у модель включається фактор, що має найбільший коефіцієнт кореляції із залежною змінною. Потім до даного рівняння добавляють фактори із найменшими коефіцієнтами кореляції доти, доки не переберуть всі фактори. Контроль адекватності моделі здійснюється за допомогою часткового *F*-критерію.

Розглянемо детальніше ***метод всіх можливих регресій***. На ***першому етапі*** цього методу встановлюють які фактори слід включати в модель. Це доволі складна процедура, яка вимагає глибокого розуміння суті модельованого процесу. З одного боку, всі фактори включені до моделі повинні мати статистичний зв'язок із результуючим показником. З іншого боку, рівняння множинної регресії адекватно відображає модельоване явище лише тоді, коли фактори є кореляційно незалежними. Якщо між факторами існує функціональний або дуже близький до нього статистичний зв'язок, то до моделі включається лише один із них, а всі інші виражаються через нього [ ].

На другому етапі здійснюється математико-статистичний аналіз факторів шляхом розрахунку парних та множинного коефіцієнтів кореляції (методика їх визначення буде наведена дещо згодом).

Рівняння лінійної багатофакторної регресії в загальному випадку має таки вигляд []:

$$\bar{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m \qquad (42)$$

За методом найменших квадратів:

$$S = \sum \left[y - (b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m) \right]^2 \to \min \tag{43}$$

Прирівнявши часткові похідні до нуля, отримаємо систему *m+1* нормальних рівнянь із *m+1* невідомими:

$$\left\{ \begin{array}{l} m b_0 + b_1 \sum x_1 + b_2 \sum x_2 + ... + b_m \sum x_m = \sum y, \\ b_0 \sum x1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + ... + b_m \sum x_1 x_m = \sum y x_1, \\ ..... \\ b_0 \sum x_m + b_1 \sum x_1 x_m + b_2 \sum x_2 x_m + ... + b_m \sum x_m^2 = \sum y x_m, \end{array} \right\} \tag{44}$$

з якої знаходимо параметри рівняння множинної лінійної регресії.

Для оцінки значимості коефіцієнтів регресії користуються *t*-критерієм:

$$t^i_{розр.} = \frac{b_i}{\sigma_{b_i}}, \tag{45}$$

де σ_{b_i} – середня квадратична похибка і-го коефіцієнта регресії.

За вибраного рівня значимості α і $V = n - m - 1$ ступенями вільності знаходимо $t_{крит.}$ Якщо $t_{розр.} > t_{крит.}$, то коефіцієнт рівняння регресії b_i слід вважати значимим і його можна використати для аналізу впливу і-тої факторної ознаки на вислідну ознаку, інакше фактор слід виключити з моделі. У такий спосіб відбувається відсіювання неістотних факторних ознак з погляду їх впливу на вислідну ознаку. Фактори, які залишились, увійдуть в модель множинної регресії.

Наступним етапом побудови кореляційно-регресійна моделі дослідження взаємозв'язку між екологічними процесами є визначення ступеня взаємозв'язку двох досліджуваних явищ (якщо ці явища взаємопоєднані) або ступеню впливу головних досліджуваних факторів на досліджуваний вплив (якщо ці зв'язки носять причинно-наслідковий характер). Найчастіше для кількісного визначення ступеня взаємозв'язку використовують ***коефіцієнт кореляції***:

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \cdot \sqrt{n \sum y^2 - (\sum y)^2}}, \tag{46}$$

Властивості коефіцієнта кореляції:

- якщо *r* набуває значення, близьке до -1, то між факторами існуєщільний обернений зв'язок;
- якщо *r* = 0, зв'язок відсутній;
- якщо *r* близьке до +1, то між факторами існує щільний прямий зв'язок;
- якщо |*r*| = 1, то між досліджуваними ознаками існує функціональнйи зв'язок;

- чим ближче значення $|r|$ до 1, тим вища щільність зв’язку між ознаками.

Перевірку значимості коефіцієнта кореляції здійснюють шляхом розрахунку величини:

$$t_{розр.} = \sqrt{\frac{r^2}{1-r^2}(n-2)}\,, \tag{47}$$

Дана величина має розподіл Стьдента для заданої ймовірності α і $V = n - 2$ ступенів свободи. Якщо $t_{розр.} > t_{крит}$, то між досліджуваними ознаками існує кореляційний зв'язок.

Для більш приблизної оцінки ступеня взаємозв'язку між досліджуваними ознаками використовують коефіцієнти Фехнера і Спірмена []. ***Коефіцієнт Фехнера*** розраховують за формулою:

$$K_{Ф} = \frac{C - H}{C + H}, \tag{48}$$

де *C*- число збігів знаків відхилень ознак х і у від їх середніх значень; *H* – число незбіжностей.

Коефіцієнт Фехнера, як і коефіцієнт кореляції, змінюється в межах від –1 до 1. При $K_{Ф} = -1$ кажуть про існування узгодженої оберненої залежності, а при $K_{Ф} = 1$ – узгодженої прямої залежності. Якщо $K_{Ф} = 0$, такої залежності не спостерігається.

Коефіцієнт Спірмена іноді ще називають коефіцієнтом рангової ко

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}, \tag{49}$$

де *n* – обсяг сукупності, а *d* – різниця рангів ознак *x* і *y*.

Для розрахунку коефіцієнта Спірмена спочатку слід побудувати з вихідних варіаційних рядів ***ранжовані*** таблиці, тобто такі таблиці, де значення варіант розташовувались би в порядку від найменшого значення до найбільшого. Потім кожне конкретне значення варіаційного ряду заміняють його рангом (порядковим номером) і рахують різницю рангів ознак *x* і *y*.

Якщо значення $\rho = -1$, то існує обернена кореляція рангів, якщо $\rho = 1$ – пряма кореляція рангів, $\rho = 0$ – кореляція рангів відсутня.