# 2 МЕТОДИ ОБРОБЛЕННЯ ДАНИХ

## 2.1 Оброблення табличних даних

***Функція*** — це правило, яке кожному елементу чи елементам (ці елементи називають аргументами) з першої множини (області визначення) ставить у відповідність один і тільки один елементі з іншої множини (значення функції або просто функція).

Спосіб задавання функції:

- Аналітичний спосіб — встановлення правил зв’язку між аргументами і функцією за допомогою формул;
- Табличний спосіб — функція задається набором окремих значень аргументів і відповідним їм значенням функції;
- Графічний спосіб — зв’язок між аргументами та функцією задається за допомогою графічної побудови.

Як вже було зазначено в попередньому розділі, сучасні обчислювальні системи в більшості є цифровими, тому вхідні та вихідні дані представляються (та й зберігаються) у вигляді окремих дискретних значень величини, тобто задаються табличним способом. Також табличні дані отримуються в процесі проведення експериментальних досліджень, як результат виміру при певних вхідних даних.

### *2.1.1 Апроксимація*

***Апроксимація*** (лат. *Approximare* — наближати) — наближене вираження одних математичних об’єктів іншими, більш простішими, наприклад, криві лінії — ламаними, ірраціональних чисел — раціональними, неперервних функцій — многочленами.

Часто при обробленні табличних даних під апроксимацією розуміють знаходження аналітичної залежності, яка наближається до цих табличних даних, не обов’язково проходить через всі точки, але описує тенденцію зміни цих параметрів. Таке наближення проводять за допомогою поліномів, а коефіцієнти поліному розраховуються методом найменших квадратів.

Поліном (многочлен) — це вираз виду:

$$Q(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$$

де $a_0, a_1, ..., a_n$ — коефіцієнти поліному;

$n$ — степінь поліному.

***Метод найменших квадратів*** — метод знаходження наближеного розв’язку надлишково-визначеної системи, який базується на пошуку мінімальної суми квадратів відхилень значення деякої функції від заданих значень:

$$\sum_{i=0}^{N}(Q(x_i)-y_i(x_i))^2 \Rightarrow \min$$

де $y_i(x_i)$ — задані значення;

$N$ — кількість значень.

Розглянемо наведений у [2] приклад пошук коефіцієнтів лінійного поліному ($n=1$) для апроксимації деякого набору табличних даних ($x$, $y$): (1, 6), (2, 5), (3, 7), (4, 10), на рис. 2.1 вони позначені колами. Необхідно знайти коефіцієнти поліному $y = a_0 + a_1 x$, так щоб побудована за даним поліномом лінія найкраще відтворювала характер поведінки даних точок. Для вирішення даного завдання складемо систему рівнянь:

$$\begin{cases} a_0 + a_1 \cdot 1 = 6 \\ a_0 + a_1 \cdot 2 = 5 \\ a_0 + a_1 \cdot 3 = 7 \\ a_0 + a_1 \cdot 4 = 10 \end{cases}$$

Тобто отримали систему з 4 рівнянь з 2 невідомими (надлишково-визначена система). Метод найменших квадратів полягає в пошуку мінімальної можливої суми квадратів похибок між лівою та правою частиною системи, тобто у пошуку мінімуму функції:

$$S(a_0,a_1) = (6-(a_0+1\cdot a_1))^2 + (5-(a_0+2\cdot a_1))^2 + \\ + (7-(a_0+3\cdot a_1))^2 + (10-(a_0+4\cdot a_1))^2$$

Мінімум визначається обчисленням часткових похідних від $S(a_0,a_1)$ по $a_0$ та $a_1$:

$$\frac{dS}{da_0} = 8\cdot a_0 + 20\cdot a_1 - 56$$

$$\frac{dS}{da_1} = 20\cdot a_0 + 60\cdot a_1 - 154$$

Прирівнявши часткові похідні до нуля отримаємо систему з двох рівнянь з двома невідомими:

$$\begin{cases} 8\cdot a_0 + 20\cdot a_1 - 56 = 0 \\ 20\cdot a_0 + 60\cdot a_1 - 154 = 0 \end{cases}$$

Розв'язавши систему отримаємо: $a_0 = 3{,}5$ та $a_1 = 1{,}4$

Рівняння прямої матиме вигляд: $y = 3{,}5 + 1{,}4 \cdot x$, рис. 2.1

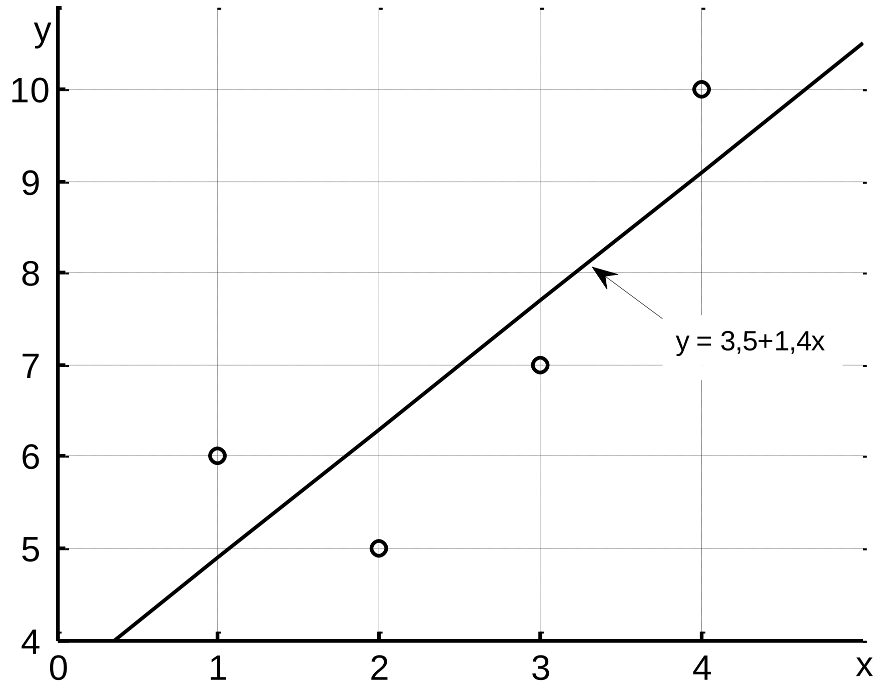


Рисунок 2.1 — Апроксимація табличних даних (кола)
за допомогою поліному 1 степеню (суцільна лінія)

З рис. 2.2 видно, що зростання степеню поліному призводить до більш якісного наближення апроксимуючої кривої до табличних даних. Тобто якість апроксимації підвищується.

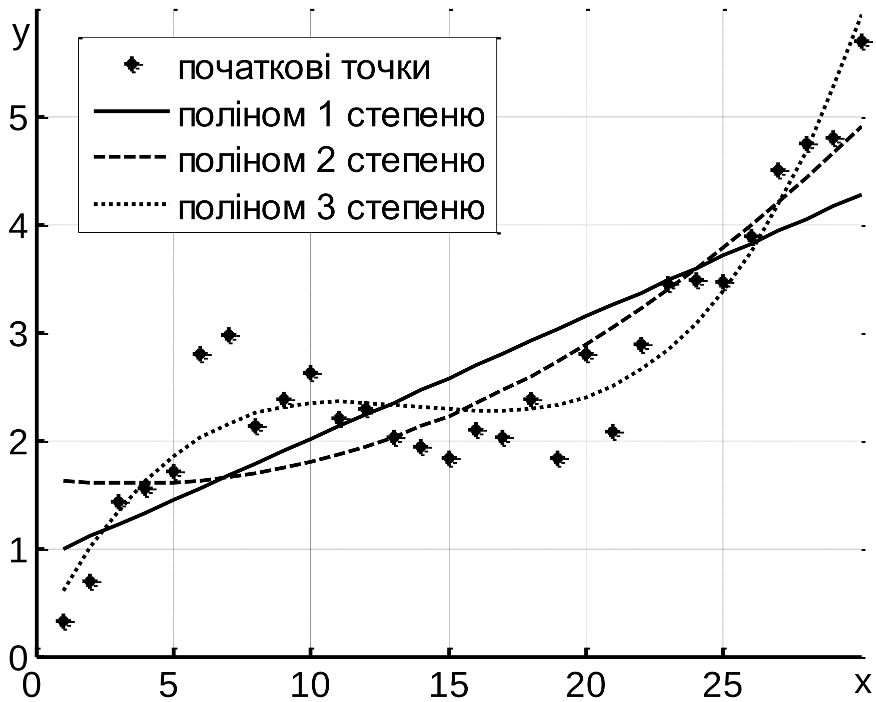


Рисунок 2.2 — Приклад апроксимації табличних даних
поліномами різного степеню

Оцінку якості апроксимації можна проводити як візуально, так і з використанням різних критеріїв придатності наближення, до яких відносяться:

– *SSE* (*Sum of squares due to error*) — сума квадратів похибок, розраховується за формулою:

$$SSE = \sum_{i=0}^{N} (y_i(x_i) - Q(x_i))^2$$

– *R–square* (*R*–квадрат) — квадрат змішаної кореляції, який розраховується за формулою:

$$Rsquare = 1 - \frac{SSE}{SST}$$

де $SST = \sum_{i=1}^{N} (y_i(x_i) - \bar{y})^2$ — повна сума квадратів;

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i(x_i)$$

– *RSME* (*Root mean Squared Error*) — середньоквадратична похибка розраховується за формулою:

$$RSME = \sqrt{\frac{SSE}{N}}$$

Зверніть увагу, що *RSME* в доповненні *cftool* (*Curve Fitting Toolbox*) *MatLAB* розраховується дещо за іншою формулою [3]:

$$RSME = \sqrt{\frac{SSE}{N - m}}$$

де $m$ — кількість параметрів моделі (при поліноміальній апроксимації $m$ відповідає кількості коефіцієнтів поліному, тобто $m = n + 1$)

Так для наведеного на рис. 2.2 випадку *SSE* для поліному 1 степеню становить 13,46, для поліному 1 степеню — 10,51, а для поліному 3 степеню — 3,726. Причому подальше зростання степеню апроксимуючого поліному не призводить до такого стрімкого зменшення похибки, так поліном 6 степеню апроксимує дані з похибкою 1,953.

### *2.1.2 Інтерполяція*

***Інтерполяція*** — спосіб знаходження проміжних значень величини за наявним дискретним набором відомих значень.

Якщо на відрізку $[a,b]$ задана функція $f(x)$, то задача інтерполяції зводиться до визначення функції $g(x)$, яка співпадає в деякому наборі точок

$\{x_1, x_1, ..., x_{n+1},\}$ з відрізку $[a,b]$(ці точки називаються вузлами інтерполяції), тобто:

$$g(x_k) = y_k,\, k = 1,2,...,n+1$$

де $\quad y_k = f(x_k)$ — відомі значення функції $f(x)$ в точках $x_k$.

Найпростішим методом інтерполяції є сполучення сусідніх відомих точок прямою лінією. Така інтерполяція отримала назву лінійна, рис. 2.3а. Також одним з відомих, але менш поширених методів інтерполяції є інтерполяція найближчим сусідом, яка полягає в тому, що проміжні значення до середини інтервалу $[x_i, x_{i+1}]$ приймають значення $x_i$, решта — $x_{i+1}$, рис. 2.3б.

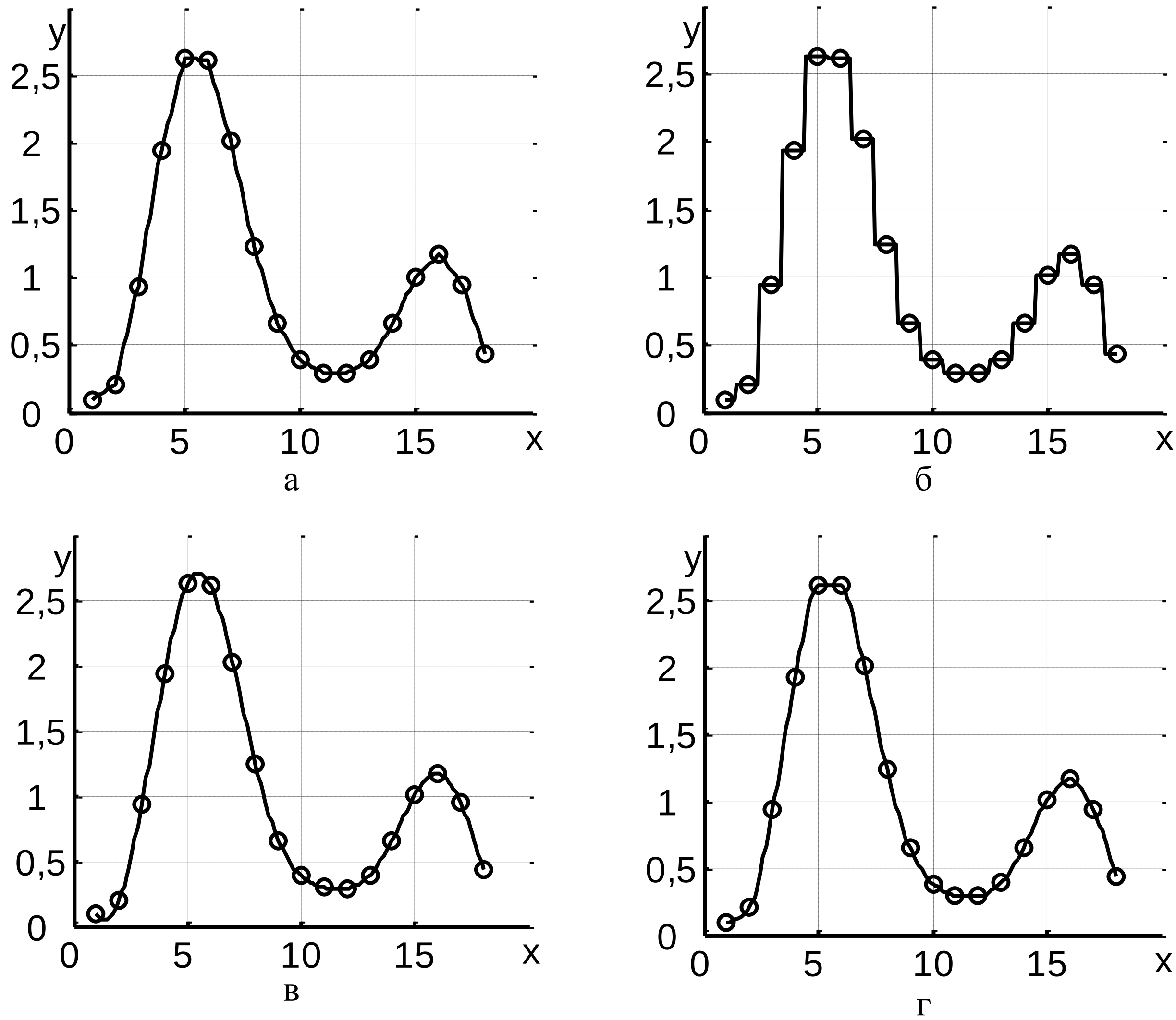


Рисунок 2.3 — Методи інтерполяції (суцільна лінія) по вузловим точкам (кола): а — лінійна; б — найближчого сусіда; в — кубічні сплайни; г — кубічний поліном Ерміта.

Більш точними, проте складнішими методами інтерполяції є методи, засновані на використанні поліномів, тобто значення між двома точками вираховується за допомогою поліному певного степеню. Зазвичай використовуються поліноми третього степеню (кубічні поліноми) [4]. Розрізняють інтерполяцію кубічними сплайнами (рис. 2.3в) та інтерполяцію кубічними поліномами Ермі-та (рис. 2.3г). В першому випадку в вузлових точках перша та друга похідна неперервна, в другому випадку неперервною є тільки перша похідна, а друга може мати розриви.

### *2.1.3 Екстраполяція*

***Екстраполяція*** — випадок інтерполяції, при якому знаходяться значення функції за межами заданого інтервалу, а не між вузловими точками.