## Лекція 6. МАТЕМАТИЧНА ОБРОБКА РЕЗУЛЬТАТІВ ЕКСПЕРИМЕНТУ

Найпоширеніший вид експерименту – це вимір якоїсь величини. Почавши вимір, ми одразу ж переконуємося, що кожен наступний вимір дасть нам трохи інший результат – ці розходження викликаються випадковими причинами різного роду: затирання в механічних частинах вимірювальних приладів і пристроїв, неточності зчитування, коливання землі (підлоги) від проїжджаючого транспорту, пориви вітру й ін. Тому говорять, що результат виміру – випадкова величина. Щоб одержати гідне довіри значення цієї випадкової величини, проводять досить багато вимірів, після чого обчислюють середньоарифметичне значення *m*, середньоквадратичне відхилення $\sigma$ й коефіцієнт варіації *v*:

$$m = \frac{\sum_{i=1}^{n} t_i}{n}; \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n} (t_i - m)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (t_i^2) - \frac{\left(\sum_{i=1}^{n} t_i\right)^2}{n}}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (t_i^2) - n \cdot m^2}{n-1}}; \qquad v = \frac{\sigma}{m},$$

де t – результат виміру; i – порядковий номер виміру; n – загальне число точок (отриманих в експерименті значень випадкової величини).

Перший варіант формули зручний для розуміння суті цього показника – характеристики відхилень випадкових значень від середнього. Другий і третій варіанти зручні для розрахунку на калькуляторі з однією коміркою пам'яті.

УВАГА! Розповсюджена помилка – уважають, що $\sum_{i=1}^{n} (t_i^2) = \left(\sum_{i=1}^{n} t_i\right)^2$. Легко упевнитися, що це не так: ряд значень t 1, 2, 3; Σt=6; (Σt)2=36; ряд квадратів цих значень $t^2$ 1, 4, 9; Σ ($t^2$)=14

Для фахівця ці три показники говорять багато про розподіл даної випадкової величини. Неспеціалістові корисно мати більше наочне представлення. Для цього звичайно будують діаграми, називані гістограма (тобто стовпчаста діаграма) і кумулята (від лат. Cumulus – купа). Порядок побудови такий.

Обчислити кількість інтервалів $z \approx 1 + 3{,}2 \lg n$ (округлити до цілого, меншого) і ширину інтервалу $h \approx (t_{max} - t_{min}) / z$, де $t_{max}$ й $t_{min}$ – відповідно, найбільше й найменше значення у вибірці (ширину інтервалу округлити до зручного значення в більшу сторону).

Гістограма – стовпчаста діаграма, що показує кількість випадкових значень, що потрапили в кожен інтервал. Наближена модель кривої щільності розподілу (тобто диференціальної кривої).

Кумулята – діаграма у вигляді ламаної лінії, що показує, скільки значень випадкової величини попадає в усі інтервали від мінус нескінченності до даного значення. Наближена модель кривої розподілу (тобто інтегральної кривої).

Приклад:

| | | | | | |
|---|---|---|---|---|---|
| 37 | 54 | 43 | **62** $t_{max}$ | 49 | 36 |
| 36 | 51 | 39 | 36 | 40 | 36 |
| **28** $t_{min}$ | 37 | 42 | 38 | 43 | 38 |
| 45 | 40 | 44 | 50 | 37 | 42 |

m = 1003 / 24 = 41.79;

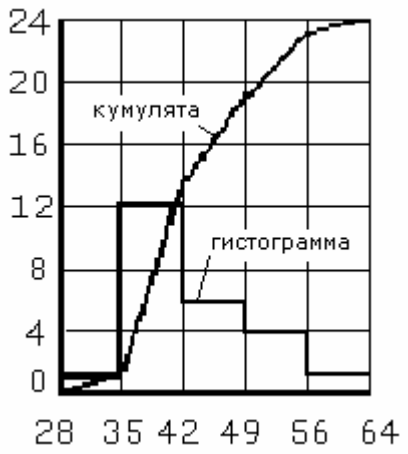
$\sigma$ = 7,271;

z = 1+3,2 lg24 = 5,42; округлюємо до 5;

h = (62 – 28) / 5 = 6,8; округлюємо до 7.

Розподіляємо значення випадкової величини по інтервалах:

| 28 – <35 | 35 – <42 | 42 – <49 | 49 – <56 | 56 – <64 |
|---|---|---|---|---|
| 28 | 37 36<br>36 39<br>36 40<br>36 37<br>38 38<br>40 37 | 43 42<br>43 45<br>44 42 | 54 49<br>51 50 | 62 |
| 1 | 12 | 6 | 4 | 1 |
| Накопичені суми | | | | |
| 1 | 13 | 19 | 23 | 24 |




Якщо збільшувати кількість вимірів, то число інтервалів буде зростати, а їхня ширина зменшуватися, і при нескінченно великій кількості вимірів ширина інтервалу стане нульовою, а діаграма зі стовпчастої перетвориться в плавну криву, яка і буде являти собою істинну криву щільності розподілу; кумулята перетвориться у плавну криву функції розподілу. Природно, ніхто не виконує нескінченно велику кількість вимірів, а для одержання передбачуваних кривих (з якимсь ступенем вірогідності) використовують різні методи апроксимації, описані в літературі й реалізовані у стандартних комп'ютерних програмах.

Ніхто не виконує нескінченно велику кількість вимірів, – а скільки їх треба виконувати? Це залежить від того, яку ми хочемо мати точність результату, тобто середньої величини: $n = (t \cdot \sigma / \varepsilon)^2$, де $t$ – показник вірогідності для даної довірчої

ймовірності $\beta$ одержуваного висновку; $\sigma$ – середньоквадратичне відхилення початкової вибірки; $\varepsilon$ – дозволена помилка вибіркової середньої. Деякі значення $t\cdot$ й $\beta\cdot$ можна брати з наступної таблички:

| $\beta$ | 0,8 | 0,85 | 0,9 | 0,95 | 0,98 | 0,99 | 0,999 |
|---|---|---|---|---|---|---|---|
| $t$ | 1,28 | 1,44 | 1,64 | 1,96 | 2,32 | 2,58 | 3,29 |

Продовжуючи початий приклад, можемо підрахувати, що якщо ми хочемо, щоб у нашу вибірку потрапили 99% всіх можливих значень, а помилка вибіркової середньої не перевищувала 1 (тобто 1,5%), тоді число вимірів повинне скласти

$n = (2{,}58\cdot 7{,}271/1)^2 = 352.$

Якщо ж ми згодні задовольнитися довірчою ймовірністю 0,9 і дозволенною помилкою 2 (3%), тоді $n = (1{,}64\cdot 7.271/2)^2 = 35{,}5 \approx 36$, тобто в 10 разів менше. Цей приклад показує, якою дорогою ціною дістається дослідникам підвищення точності результатів експерименту.

Методи математичної обробки результатів експерименту розроблені дуже докладно й описані в багатьох книгах.

## АПРОКСИМАЦІЯ ЕКСПЕРИМЕНТАЛЬНИХ ЗАЛЕЖНОСТЕЙ

Апроксимація (від лат. approximare – наближатися) – це підбір математичного виразу (формули), що щонайкраще наближується до отриманої в експерименті залежності показника $y$ від фактора $x$ (часто цю процедуру називають також "вирівнювання" й "згладжування"). Дослідникові доводиться займатися цим постійно: звичайно, можна ввести в комп'ютер результати експерименту в табличній формі й скласти програму приблизного обчислення $y$ при проміжних значеннях $x$ (це називається інтерполяція) або за межами використаного в експерименті діапазону значень x (це називається екстраполяція; зауважимо, що отримані шляхом інтерполяції значення досить надійні, їм можна довіряти, а от до результатів екстраполяції слід ставитися дуже обережно).

Але набагато зручніше знайти вираз (формулу, функцію), які можна підставляти в якісь математичні моделі в загальному виді. Таких виразів можна записати безліч. Як же оцінити, який з них буде наближатися найкраще? А який критерій якості наближення? Найменше відхилення від важливої для нас точки або групи точок? А як бути з іншими точками? Або такий критерій – сума відхилень по

всіх точках? Але може трапитися, що позитивні й негативні відхилення взаємно знищаться й дадуть у сумі нуль, хоча наближення буде дуже поганим.

**Метод найменших квадратів (МНК).** Карл Фрідріх Гаусс (1777-1855) запропонував як такий критерій використати суму квадратів відхилень у всіх точках. Із декількох формул та, що дасть найменше значення цієї суми, і буде найкращою. Критеріальне рівняння МНК:

$$\sum_{i=1}^{n}\left[y_i - f\left(x_i, A, B, C...\right)\right]^2 \Rightarrow \min;$$

де $i$ – порядковий номер експериментальної точки $y_i(x_i)$; $n$ – кількість прийнятих у розрахунок експериментальних точок; ; $f(x_i,A,B,C\ldots)$ – розрахункова формула, за допомогою якої ми намагаємося апроксимувати результати експерименту; A,B,C… – чисельні коефіцієнти в цій формулі.

Виходить, ми намагаємося знайти мінімум (тобто екстремум) функції Σ... від... від якого аргументу? Від x? Ні. Всі значення ***x*** и ***y***, які ми одержали в експерименті, це вже константи. Ми не можемо варіювати ними як заманеться. У нашій владі тільки загальний вигляд функції та її коефіцієнти – саме їх ми й намагаємося знайти. Виходить, ми шукаємо мінімум функції Σ… залежно від коефіцієнтів A, B, C… Як відомо з математики, у точках екстремуму перша похідна функції дорівнює нулю. А оскільки в нас кілька аргументів (A,B,C...), критерієм буде рівність нулю частинних похідних $\frac{\partial\sigma}{\partial a}=0;\frac{\partial\sigma}{\partial b}=0;\frac{\partial\sigma}{\partial c}=0;$ і т.д. Розглянемо шлях рішення задачі пошуку коефіцієнтів A, B, C наприклад, для полінома другого ступеня (квадратного тричлена) виду $y = Ax^2 + Bx + C.$

Переписуємо критеріальне рівняння для цього випадку:

$$\sum_{i=1}^{n}\left[y_i - Ax^2 - Bx - C\right]^2 \Rightarrow \min;$$

Перша частинна похідна (по A):

$$\frac{\partial\sigma}{\partial a} = 2\sum_{i=1}^{n}\left[y_i - Ax_i^2 - Bx_i - C\right]\cdot\left(-x_i^2\right) = 0;$$

$$\sum_{i=1}^{n}\left[y_i - Ax_i^2 - Bx_i - C\right]\cdot\left(x_i^2\right) = \sum_{i=1}^{n}y_i x_i^2 - A\sum_{i=1}^{n}x_i^4 - B\sum_{i=1}^{n}x_i^3 - C\sum_{i=1}^{n}x_i^2 = 0;$$

або, якщо переписати останнє рівняння в канонічному вигляді,

$$A\sum_{i=1}^{n}x_i^4+B\sum_{i=1}^{n}x_i^3+C\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}y_ix_i^2\ .$$

Аналогічно знаходимо похідні по b й c:

$$\frac{\partial\sigma}{\partial b}=2\sum_{i=1}^{n}\left[y_i-Ax_i^2-Bx_i-C\right]\cdot\left(-x_i\right)=0;$$

$$\sum_{i=1}^{n}\left[y_i-Ax_i^2-Bx_i-C\right]\cdot\left(x_i\right)=\sum_{i=1}^{n}y_ix_i\ -A\sum_{i=1}^{n}x_i^3-B\sum_{i=1}^{n}x_i^2-C\sum_{i=1}^{n}x_i\ =0;$$

$$A\sum_{i=1}^{n}x_i^3+B\sum_{i=1}^{n}x_i^2+C\sum_{i=1}^{n}x_i\ =\sum_{i=1}^{n}y_ix_i$$

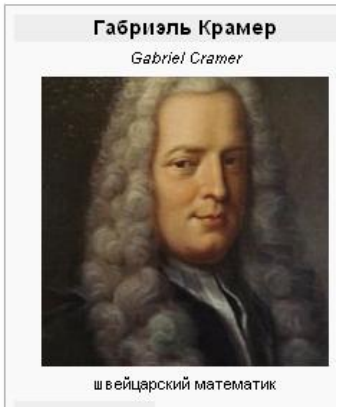
$$\frac{\partial\sigma}{\partial c}=2\sum_{i=1}^{n}\left[y_i-Ax_i^2-Bx_i-C\right]\cdot\left(-1\right)=0;$$

$$\sum_{i=1}^{n}\left[y_i-Ax_i^2-Bx_i-C\right]=\sum_{i=1}^{n}y_i-A\sum_{i=1}^{n}x_i^2-B\sum_{i=1}^{n}x_i\ -C\sum_{i=1}^{n}1=0;$$

$$A\sum_{i=1}^{n}x_i^2+B\sum_{i=1}^{n}x_i\ +C\cdot n=\sum_{i=1}^{n}y_i$$

Отже, отримана система із трьох рівнянь із трьома невідомими A, B й C.

$$\begin{cases}A\sum_{i=1}^{n}x_i^4+B\sum_{i=1}^{n}x_i^3+C\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}y_ix_i^2\\A\sum_{i=1}^{n}x_i^3+B\sum_{i=1}^{n}x_i^2+C\sum_{i=1}^{n}x_i\ =\sum_{i=1}^{n}y_ix_i\\A\sum_{i=1}^{n}x_i^2+B\sum_{i=1}^{n}x_i\ +C\cdot n=\sum_{i=1}^{n}y_i\end{cases}$$




Будемо вирішувати цю систему по методу Крамера:

$$A=\frac{\Delta A}{\Delta};B=\frac{\Delta B}{\Delta};C=\frac{\Delta C}{\Delta};$$

Тут $\Delta$ – головний визначник системи, складений з коефіцієнтів при невідомих.

$$\Delta=\begin{vmatrix}\sum x^4 & \sum x^3 & \sum x^2\\ \sum x^3 & \sum x^2 & \sum x\\ \sum x^2 & \sum x & n\end{vmatrix}=$$

$$=\sum x^4\left(n\sum x^2-\left(\sum x\right)^2\right)-\sum x^3\left(n\sum x^3-\sum x\cdot\sum x^2\right)+\sum x^2\left(\sum x\cdot\sum x^3-\left(\sum x^2\right)^2\right)$$

Подальші перетворення в загальному виді не мають змісту. При обчисленнях варто записати й зберегти значення виразів у круглих дужках – вони будуть використані повторно. Визначники ΔA, ΔB й ΔC виходять шляхом заміни стовпця коефіцієнтів при невідомому A, B або C відповідно на стовпець вільних членів.

Дуже часто апроксимацію виконують лінійною функцією виду $y = Ax + B$ Коефіцієнти A й B знаходять аналогічним способом, тільки формули для них набагато простіше й можуть бути записані в загальному виді.

$$\sum_{i=1}^{n}\left[y_i - Ax_i - B\right]^2 \Rightarrow \min; \qquad \frac{\partial\sigma}{\partial a} = 0; \frac{\partial\sigma}{\partial b} = 0;$$

$$\frac{\partial\sigma}{\partial a} = 2\sum_{i=1}^{n}\left[y_i - Ax_i - B\right]\cdot\left(-x_i\right) = 0;$$

$$\sum_{i=1}^{n}\left[y_i - Ax_i - B\right]\cdot\left(x_i\right) = \sum_{i=1}^{n} y_i x_i - A\sum_{i=1}^{n} x_i^2 - B\sum_{i=1}^{n} x_i = 0;$$

$$A\sum_{i=1}^{n} x_i^2 + B\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i x_i$$

$$\frac{\partial\sigma}{\partial b} = 2\sum_{i=1}^{n}\left[y_i - Ax_i - B\right]\cdot\left(-1\right) = 0;$$

$$\sum_{i=1}^{n}\left[y_i - Ax_i - B\right] = \sum_{i=1}^{n} y_i - A\sum_{i=1}^{n} x_i - B\sum_{i=1}^{n} 1 = 0;$$

$$A\sum_{i=1}^{n} x_i + B\cdot n = \sum_{i=1}^{n} y_i$$

$$\begin{cases} A\sum_{i=1}^{n} x_i^2 + B\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i x_i \\ A\sum_{i=1}^{n} x_i + B\cdot n = \sum_{i=1}^{n} y_i \end{cases} \qquad A = \frac{\Delta A}{\Delta}; B = \frac{\Delta B}{\Delta};$$

$$\Delta = \begin{vmatrix} \sum x^2 & \sum x \\ \sum x & n \end{vmatrix} = n\sum x^2 - \left(\sum x\right)^2;$$

$$\Delta A = \begin{vmatrix} \sum yx & \sum x \\ \sum y & n \end{vmatrix} = n\sum yx - \sum y\sum x; A = \frac{n\sum yx - \sum y\sum x}{n\sum x^2 - \left(\sum x\right)^2};$$

$$\Delta B = \begin{vmatrix} \sum x^2 & \sum yx \\ \sum x & \sum y \end{vmatrix} = \sum y\sum x^2 - \sum yx\sum x; B = \frac{\sum y\sum x^2 - \sum yx\sum x}{n\sum x^2 - \left(\sum x\right)^2}.$$

Отримана формула для B трохи громіздка. Із другого рівняння системи можна одержати більше компактну формулу: $B = \frac{\sum y - A\sum x}{n}$. Недолік цієї формули в тім, що якщо коефіцієнт A обчислений з помилкою, тоді й коефіцієнт B буде невірний. При виникненні сумнівів його можна перевірити по більше громіздкій, зате незалежній формулі.

Для апроксимації нелінійних залежностей широко використовують степеневу, показникову й логарифмічну функції. Загальний вигляд степеневої функції – $y = D \cdot x^E$. Її можна привести до лінійного виду, перетворивши за допомогою логарифмування: $\log y = \log D + E \cdot \log x = E \cdot \log x + \log D.$ Тепер можна використати вже готові формули для коефіцієнтів лінійної функції, якщо внести відповідні заміни:

$$E = \frac{n\sum(\log y \cdot \log x) - \sum \log y \sum \log x}{n\sum \log^2 x - \left(\sum \log x\right)^2}.$$

$\log D = \frac{\sum \log y \sum \log^2 x - \sum(\log y \cdot \log x)\sum \log x}{n\sum \log^2 x - \left(\sum \log x\right)^2}$. або $\log D = \frac{\sum \log y - E\sum \log x}{n}$.

Так само одержимо коефіцієнти для показникової функції виду $y = F \cdot G^x$. Після логарифмування $\log y = \log F + x \cdot \log G = \log G \cdot x + \log F.$

$$\log G = \frac{n\sum(x \cdot \log y) - \sum \log y \sum x}{n\sum x^2 - \left(\sum x\right)^2}. \quad \log F = \frac{\sum \log y \sum x^2 - \sum(x \cdot \log y)\sum x}{n\sum x^2 - \left(\sum x\right)^2}.$$

$\log F = \frac{\sum \log y - \log G \sum x}{n}$.

Логарифмічна функція $y = K \cdot \log x + M.$ лінійна відносно аргументу log x, її коефіцієнти $K = \frac{n\sum(y \cdot \log x) - \sum y \sum \log x}{n\sum \log^2 x - \left(\sum \log x\right)^2}$.

$M = \frac{\sum y \sum \log^2 x - \sum(y \cdot \log x)\sum \log x}{n\sum \log^2 x - \left(\sum \log x\right)^2}$. $M = \frac{\sum y - K\sum \log x}{n}$.

Взагалі апроксимацію можна виконувати на комп'ютері в пакеті Excel, використовуючи опцію "додати лінію тренда" у закладці "Діаграми". Однак відзначалися випадки, коли деякі версії Excel дають гіршу апроксимацію, чим

рахунок за виведеними тут формулами (ручний або в пакеті Mathcad). Додамо, що для апроксимації можна використати будь-які функції, що дають потрібний вигляд кривої, наприклад, тригонометричні: синусоїда гарна при апроксимації картини коливальних процесів; сума двох котангенсоїд з різними фазами й масштабними коефіцієнтами була успішно використана, наприклад, при апроксимації експериментальної залежності коефіцієнта тертя в гальмівній парі від температури. Експоненційна функція $e^x$ добре описує, наприклад, загасання коливань.

МНК дозволяє знайти коефіцієнти до обраної вами апроксимувальної функції, але нічого не говорить про те, чи гарна ця функція взагалі. Щоб вирішити, яка з декількох функцій краще описує експериментальні дані, підраховують основну помилку апроксимації $\sigma_o = \sqrt{\frac{\sum [y_i - y(x_i)]^2}{n-1}}$, де $y_i$ – отримані в експерименті значення $y$; $y(x_i)$ – отримані розрахунком за формулою апроксимації значення $y$ при відповідних значеннях $x_i$. Іноді зручніше оцінювати апроксимацію за відносною помилкою $\bar{\varepsilon} = \sqrt{\frac{\sum [1 - y(x_i)/y_i]^2}{n-1}} 100\%$. Очевидно, наближення тим вдаліше, чим менше помилка апроксимації.

Кореляція (від пізнелат. correlatio – співвідношення) термін, застосовуваний у різних галузях науки й техніки для позначення взаємозалежності, взаємної відповідності, співвідношення понять, підприємств, предметів, функцій.

**Кореляційний аналіз** використовують для того, щоб установити факт наявності зв'язку між двома параметрами x и y. У найпростішому випадку обчислюють коефіцієнт кореляції: $r = \frac{n\sum xy - \sum x \cdot \sum y}{\sqrt{\left[n\sum\left(x^2\right) - \left(\sum x\right)^2\right] \cdot \left[n\sum\left(y^2\right) - \left(\sum y\right)^2\right]}}$.

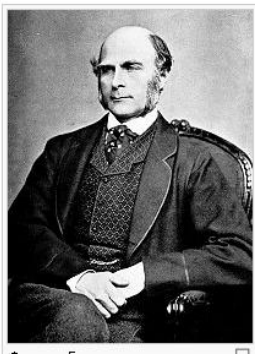
Якщо r = 0,9...1, це говорить про функціональну залежність; 0,7...0,9 – сильний ступінь кореляції; 0,5...0,7 – помітний ступінь кореляції; 0,3...0,5 – помірний ступінь кореляції; 0,1...0,3 – слабкий ступінь кореляції; 0...0,1 – відсутність кореляції. Негативні значення r відповідають випадку, коли y убуває зі збільшенням x.

Якщо зв'язок між x и y не лінійний, а більше складний, коефіцієнт кореляції непридатний для оцінки цього зв'язку; у таких випадках використовують

кореляційне відношення $R = \sqrt{1 - \frac{\sigma_x^2}{\sigma_y^2}}$, де $\sigma_x^2 = \sigma_o^2 = \frac{\sum [y_i - y(x_i)]^2}{n-1}$; $\sigma_y^2 = \frac{\sum [y_i - \bar{y}]^2}{n-1}$;

неважко зрозуміти, що при перевірці гіпотез про придатність різних функцій для опису цього зв'язку значення кореляційного відношення різні.

Програма Excel виводить як показник вірогідності апроксимації квадрат кореляційного відношення, тобто підкореневий вираз.



Френсіс Гальтон (Galton). Двоюрідний брат Чарльза Дарвіна. Мандрівник, метеоролог, біолог. Статистика в біології. Дерматогліфіка (наука про відбитки пальців). Уперше запропонував формулу для обчислення коефіцієнта кореляції