Але гігантські потоки даних нічого не варті, якщо ми не в стані провести їх ефективно використати. Саме тут нам у нагоді можуть стати математичні методи обробки та аналізу інформації.

Основні етапи розв`язання задачі **обробки та аналізу даних**:

1. отримання даних,
2. обробка даних,
3. аналіз результатів обробки даних.

Детальніше:

1 етап: потрібно забезпечити збір та збереження.
Високоточні прилади для отримання інформації.
Носії інформації: HDD, SSD, USB flash-drive і т.п.

! 2 етап: використовуємо широкий спектр ППП.

! 3 етап: Головне дати коректну інтерпретацію результатів обробки даних та зробити правильні висновки з них.

Вам відомий вислів:

*«**Хто володіє інформацією, той володіє світом**, якщо вміє її професійно обробити та зробити правильні висновки з результатів обробки».*

Основні розділи курсу по обробці та аналізу інформації (даних):

- попередня обробка даних,
- кореляційний аналіз,
- регресійний аналіз,
- дисперсійний аналіз,
- коваріаційний аналіз,
- аналіз часових рядів,
- дискримінантний аналіз,
- кластерний аналіз і т.д.

Інформація (дані), як правило, **отримується** при наявності помилок вимірювання, збурень або якихось інших непередбачуваних впливів, тобто **в умовах невизначеності**.

Підходи, які використовуються **для опису об'єктів в умовах невизначеності:**

- ймовірнісний підхід (найчастіше),
- теорія розмитих множин по Заде,
- мінімаксний (гарантований) підхід (set membership approach) (автор: професор Микола Федорович Кириченко (15.06.1940-19.12.2008)),
- ???.

Якщо порівнювати **реальність**, яка нас оточує, з **діамантом**, то кожен з підходів для опису об'єктів в умовах невизначеності віддзеркалює тільки одну з його граней, а поява нових таких підходів дозволить віддзеркалити інші грані цього діаманту. Таким чином, поступово здійснюється пізнання реальності, яка нас оточує.

Все нове, що виникає в методах по обробці та аналізу даних з'являється як відгук на деяку потребу на практиці, яка не вкладається у рамки вже відомих постановок задач. Тобто у процесі подальшого розвитку арсеналу методів відбувається **повний диктат практики**.

Класифікація змінних

Інформація (дані) – це результат проведення спостережень над деякими змінними. Дамо їх класифікацію.

Змінні поділяються на ***скалярні*** змінні та ***векторні*** змінні.

Так як значення скалярних змінних, які спостерігаються, можуть бути як кількісні так і якісні, то це дозволяє зробити їх поділ на ***кількісні*** та ***якісні***.

Означення. ***Кількісною скалярною змінною*** називається скалярна змінна, область значень якої є поле дійсних чисел ($\mathbb{R}$).

Приклади: час, температура, відстань, площа, об'єм, швидкість, прискорення, вага, маса, тиск, і т.п.

Означення. ***Якісною скалярною змінною*** називається скалярна змінна, область значень якої є деяка множина, яка не є підмножиною поля дійсних чисел. Елементи цієї множини називають ***градаціями*** (категоріями).

Якісні скалярні змінні поділяють на:

- ***ординальні (порядкові),***
- ***номінальні (класифікаційні).***

Означення. Якісна скалярна змінна називається ***ординальною***, якщо на множині її градацій *задано природній порядок*, інакше вона називається ***номінальною***.

Приклади:

- *номінальної змінної*: змінна «Навчальна дисципліна», яка приймає в якості своїх значень назви дисциплін, що вивчаються студентами факультету: математичний аналіз, алгебра, програмування, теорія ймовірностей і математична статистика, диференційні рівняння і т.д. Тут градації не впорядковані.

- *ординальної змінної*: змінна «Військове офіцерське звання», яка приймає в якості своїх значень військові офіцерські звання - молодший лейтенант, лейтенант, старший лейтенант, капітан (капітан-лейтенант для ВМС), майор (капітан ІІІ рангу для ВМС), підполковник (капітан ІІ рангу для ВМС), полковник (капітан І рангу для ВМС), генерал-майор (контр-адмірал для ВМС), генерал-лейтенант (віце-адмірал для ВМС), генерал-полковник (адмірал для ВМС), генерал армії (адмірал флоту для ВМС). Ці градації строго впорядковані.

Крім цього якісні скалярні змінні поділяють на *категорізовані* та *некатегорізовані*.

Означення. Якісна скалярна змінна називається ***категорізованою***, якщо для неї повністю визначена множина градацій та правило *однозначного* віднесення довільного її значення до певної градації. У протилежному випадку якісну скалярну змінну називають ***некатегорізованою***.

Приклади:

- *категорізованої змінної*: змінна «Військове офіцерське звання»,
- *некатегорізованої змінної*:
  - змінна «Навчальна дисципліна»,
  - змінна «Темперамент»: (флегматик, меланхолік, сангвінік, холерик).

Скалярні змінні також поділяють на ***дискретні*** та ***неперервні***.

Зауваження. Надалі будемо використовувати наступні позначення для характеристик випадкової величини ξ:

$F_\xi(x) = P\{\xi < x\}$ – функція розподілу,

$p_\xi(x)$ – функція щільності,

$\{y_i, p_i\}_{i=1}^m$ – полігон ймовірностей, коли ξ дискретна в.в., яка набуває значення y_i з ймовірностями p_i, $i = \overline{1,m}$.

Та їх вибіркові аналоги:

$\hat{F}_\xi(x)$ – емпірична (вибіркова) функція розподілу,

$\hat{p}_\xi(x)$ – емпірична (вибіркова) функція щільності,

$\{y_i, \hat{p}_i\}_{i=1}^m$ – полігон частот (відносних), коли ξ - дискретна в.в.

Групування даних

Труднощі при обробці вибірок великого. Підходи їх подолання:

- починаємо обробляти вибірку, а що не встигнемо *залишається на обробку наступникам* (ситуація роботи з надзвичайно цінною інформацією),
- *обробляємо тільки частину оригінальної вибірки*, яка з нашої точки зору є найбільш інформативною та представницькою,
- спочатку в оригінальній вибірці кожну групу значень близьких між собою замінюють найбільш характерним значенням з деяким показником, а далі вже *обробляють тільки вже побудовану множину пар найбільш характерних значень зі своїми показниками*, причому об`єм новоутвореної вибірки вибирається реальним для обробки за потрібний час.

Один з варіантів реалізації останнього підходу: ***групування даних***.

Групування даних, детальніше його оригінальний алгоритм:

I. *Випадок обробки скалярних спостережень.* Нехай

$$\xi: x_1, x_2, \ldots, x_n.$$

Проведемо ***групування даних за скалярною змінною*** ξ.

Як правило, його застосовують при обробці спостережень над неперервними змінними, коли об'єм вибірки $n > 50$, а над дискретними змінними, коли кількість її значень $m > 10$.

Визначаємо $x_{\min} = \min\limits_i x_i$ та $x_{\max} = \max\limits_i x_i$. та розбиваємо інтервал

$$[x_{\min}, x_{\max}]$$

на s однакових підінтервалів $[c_{i-1}, c_i)$, $i = \overline{1,s}$. А $x_{\max}$ в $[c_{s-1}, c_s]$.

Ці підінтервали називають ***інтервалами групування***.

Вибір s: $5 \le s \le 30$.

Формула Стерджеса: $s = 1 + [\log_2 n]$.

Для кожного з підінтервалів $[c_{i-1}, c_i)$ підраховують значення його центральної точки:

$$x_i^* = \frac{c_{i-1} + c_i}{2} \quad \text{та} \quad n_i,$$

де n_i - кількість вимірів з вибірки, які потрапили в i – ий підінтервал $[c_{i-1}, c_i)$. $\left(i = \overline{1,s}\right)$.

В результаті, здійснено перехід від оригінальної вибірки

$$x_1, x_2, \ldots, x_n \qquad \text{до} \qquad \left\{x_i^*, n_i\right\}_{i=1}^{s}, \quad \left(n = \sum_{i=1}^{s} n_i\right).$$

Рекомендується вибирати $n_i \geq 5,\ i = \overline{1,s}$.

Зауваження. Про вибір інтервалів групування не однакової довжини на практиці. …

Проведемо ***групування даних за векторною змінною*** ζ.

Спочатку проведемо групування даних за кожною скалярною змінною ξ_i, як це було описано раніше $\left(i = \overline{1,q}\right)$. Нехай в результаті область значень скалярної змінної ξ_i розбилася на s_i інтервалів групування $\left(i = \overline{1,q}\right)$.

Тоді, в свою чергу, область значень вектора $\vec{\zeta}$ розіб'ється на $s = \prod_{i=1}^{q} s_i$ гіперпаралелепіпедів. Таким чином, у векторному випадку вже працюють не з інтервалами групування, а з їх аналогами ***гіперпаралелепіпедами групування***.

Після цього, аналогічно скалярному випадку, для $i-$ого гіперпаралелепіпеда групування визначають значення його центральної точки $\vec{y}_i^*$ та відповідну кількість вимірів n_i з вибірки $\{\vec{y}_i\}_{i=1}^n$, яка потрапили у цей $i-$ий гіперпаралелепіпед групування $\left(i=\overline{1,s}\right)$. У підсумку, здійснили перехід від оригінальної вибірки

$\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_n$ до набору значень пар $\left\{\vec{y}_i^*, n_i\right\}_{i=1}^s$, $\left(\sum_{i=1}^s n_i = n\right)$.

Самостійна робота №1. З навчального посібника «Слабоспицький О.С. Аналіз даних. Попередня обробка, 2001». Пропрацювати матеріал наведений у **розділі 1.3**:
Датчики псевдовипадкових чисел та моделювання дискретних та неперервних випадкових величин (пропустити десь 3 сторінки).