

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І  
ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ**

**ОБРОБКА ІНФОРМАЦІЇ В КОМП'ЮТЕРНО-ІНТЕГРОВАНИХ  
СИСТЕМАХ АВТОМАТИЗАЦІЇ**

**Конспект лекцій**

для здобувачів третього (освітньо-наукового) рівня вищої освіти за  
спеціальністю 151 «Автоматизація та комп'ютерно-інтегровані технології»  
галузі знань 15 «Автоматизація та приладобудування»  
(кваліфікація: PhD доктор філософії)

Київ-2022

**УДК 621.391(075.8)**

Конспект лекцій з курсу «Обробка інформації в комп'ютерно-інтегрованих системах автоматизації» призначений для здобувачів третього (освітньо-наукового) рівня вищої освіти за спеціальністю 151 «Автоматизація та комп'ютерно-інтегровані технології» галузі знань 15 «Автоматизація та приладобудування» (кваліфікація: PhD доктор філософії).

Рекомендовано до друку рішенням вченої ради ННІ енергетики, автоматики і енергозбереження Національного університету біоресурсів і природокористування України (протокол № 11 від 17 грудня 2021 р.).

**НАВЧАЛЬНЕ ВИДАННЯ**  
**ОБРОБКА ІНФОРМАЦІЇ В КОМП'ЮТЕРНО-ІНТЕГРОВАНИХ**  
**СИСТЕМАХ АВТОМАТИЗАЦІЇ**

**Конспект лекцій**

Укладачі: докт. техн. наук, професор В.В. Коваль,  
канд. техн. наук, доцент М.О. Кіктєв.

Рецензенти:

д.т.н., проф. Горобець В.Г.  
д.т.н., проф. Шворов С.А.

Технічний редактор В.М. Поліщук  
Відповідальний за випуск С.В. Драгнєв  
Комп'ютерна верстка С.В. Драгнєв

Підписано до друку 17.12.21  
Ум. друк. арк. 6,9  
Наклад 20 прим.

Формат 60x84/16  
Обл. вид. арк. 6,2  
Зам. № 2919

Видавничий центр НУБіП  
03041, вул. Героїв Оборони, 15

## ЗМІСТ

|                                                                                                                              |    |
|------------------------------------------------------------------------------------------------------------------------------|----|
| ВСТУП .....                                                                                                                  | 4  |
| Лекція № 1. Методи попереднього аналізу та зниження розмірностей багатовимірних даних.....                                   | 5  |
| Лекція № 2. Методи обробки багатовимірних даних у OLAP-системі..                                                             | 21 |
| Лекція № 3. Класифікація даних з використанням детермінованих моделей .....                                                  | 31 |
| Лекція № 4. Методи статистичного оцінювання і порівняння вибірок. Непараметричні методи перевірки однорідності вибірок... .. | 46 |
| Лекція № 5. Дисперсійний аналіз. Методи обробки рангових даних.....                                                          | 60 |
| СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ.....                                                                                        | 71 |

## ВСТУП

Лекційний матеріал складено відповідно до чинного силабусу навчальної дисципліни «Обробка інформації в комп'ютерно-інтегрованих системах автоматизації» для здобувачів ступеня доктора філософії, які навчаються за спеціальністю 151 «Автоматизація та комп'ютерно-інтегровані технології» та освітньою програмою «Автоматизація та комп'ютерно-інтегровані технології». Представлений лекційний матеріал, необхідний для розуміння основних термінів та понять у інформаційній сфері та для виконання конкретних практичних завдань з обробки інформації, а також для самоперевірки отриманих знань.

Виконання практичних завдань, пов'язаних з освоєнням принципів статистичного аналізу даних, базових засад аналізу багатовимірних даних, сприятиме закріпленню, поглибленню та узагальненню теоретичних основ курсу, а також розвитку навичок самостійної творчої роботи здобувачів у процесі їх навчання. Мета даного навчального видання – допомогти здобувачам отримати необхідні знання та уміння з теорії, методик та практики проведення обробки даних, а також методології аналізу результатів при вирішенні задач обробки великих масивів даних сучасними методами.

## **Лекція № 1**

### **Методи попереднього аналізу та зниження розмірностей багатовимірних даних**

#### **1. Визначення інформаційної технології. Поняття інформаційної технології**

У минулому термін «технологія» широко використовувався лише при описі певних виробничих процесів. Проте на сьогоднішній день існує більш широкий спектр застосувань даного терміну, тож розглянемо деякі з них.

«Технологія – це сукупність методів виготовлення, обробки, зміни властивостей, стану, форми сировини, матеріалу або напівфабрикату, здійснюваних у процесі виробництва продукції» [1].

«Технологія – це сукупність виробничих методів і процесів галузі виробництва, а також науковий опис способів виробництва ...».

«Технологія - це: 1) сукупність методів виготовлення, обробки, зміни властивостей, стану, форми сировини, матеріалу або напівфабрикату в професії виробництва, наприклад, технологія металів, хімічна технологія, технологія будівельних робіт; 2) наука про способи впливу на матеріали, сировину або напівфабрикати відповідними знаряддями виробництва».

Стає очевидним, що технологія – це методологія, яка дозволяє знайти наукове вирішення для практичних завдань. Відповідно розквіт сучасної технології можна безпосередньо пов'язати з науково-технічною революцією [1].

Усі наведені вище визначення орієнтовані здебільшого на промисловість. Однак, термін «технологія» з часом набув нових значень для інших сфер діяльності людини, наприклад, педагогіки, соціології, культури, програмування, тощо.

Отже, на сьогоднішній день термін «технологія» можна розглядати по-різному. З точки зору філософії технологія є вченням про оптимізацію діяльності. Для міждисциплінарного рівня визначення технології може бути наступним: процес, який визначається сукупністю методів та засобів обробки, виготовлення, зміни властивостей, форми, стану сировини чи матеріалу. У більш загальному технологію можна визначити як набір методів та засобів, що використовуються для оптимального застосування та перетворення енергії, матеріалів та інформації в інтересах суспільства, окремих людей та навколишнього середовища.

Заміна матеріального об'єкта на інформаційний дозволяє використовувати термін «технологія» в області, що стосується обробки інформації із застосуванням сучасних засобів автоматизації та комп'ютерних технологій.

Інформація є одним з найцінніших ресурсів сучасного суспільства поряд з такими традиційними матеріальними ресурсами як газ, нафта, корисні копалини, тощо. Термін «інформація» з'явився порівняно недавно.

Перші фундаментальні наукові праці в області теорії інформації з'явилися близько сімдесяти років тому. Трохи пізніше інформація була віднесена до найважливіших наукових понять, а також названа першоосновою перспективних нових галузей науки і техніки.

Актуальним для умов сучасного ринку є визначення терміну «інформація»: «Інформація – це ті послуги або продукти, які призначені для передачі знань у максимально доступній для потенційного кінцевого споживача формі їх виробником».

Поняття інформації в якомусь сенсі є більш широким, але більш вузьким ніж знання. Загальний потік інформації, який через органи чуття надходить із зовнішнього світу в мозок людини, виражається числом більше 100 000 бітів в секунду. Проте лише незначна частина цієї інформації розпізнається свідомістю людини. Ця інформація в своїй результативній формі може виступати у вигляді знання. «... Знання є протилежним незнанню, тобто відсутності перевіреної інформації про що-небудь ... ». Знання можуть з'явитися тільки після отримання і переробки інформації.

Проте для технічних наук, як, наприклад, інформатика чи програмування, визначення терміну «інформація» не може ґрунтуватися на такому понятті, як знання, а також не може спиратися виключно на об'єктивність свідчень і фактів. При використанні автоматичної або автоматизованої обчислювальної техніки, яка може обробляти інформацію без участі людини (або при незначній участі) не може йти мова про знання. Ці засоби можуть працювати з абстрактною, штучною і навіть помилковою інформацією, яка не має адекватного чи об'єктивного відображення ні в суспільстві, ні в природі. Можна зробити висновок, що інформація являється продуктом певної взаємодії між даними й адекватними для них методами.

Найчастіше інформатика розглядає інформацію як поняття, дані чи відомості, пов'язані між собою концептуально, що змінюють наші уявлення про об'єкти чи явища навколишнього світу. Крім терміну «інформація» в інформатиці часто використовується термін «дані». Дані можуть розглядатися як записані спостереження або ознаки, які не використовуються, а тільки зберігаються. Дані можуть стати інформацією у тому випадку, коли вони беруть участь у процесі зняття невизначеності. Отже, можна стверджувати, що дані, які використовуються, є інформацією.

Ідея включення інформації в ланцюг виробництва інформаційного об'єкта належить В.М. Глушкову. Він у 1982 році в своїй книзі «Основи безпаперової інформатики» дав наступне визначення: «Інформаційні технології – це процеси, в яких основною продукцією, яка переробляється, є інформація». Звідси можна зробити висновок про те, що інформаційні технології (ІТ) використовувалися завжди, бо завдання обробки, поширення та накопичення інформації стояли перед людством на всіх етапах його розвитку.

Існує думка, що методичні системи раніше не називали ІТ тому, що появу цього терміну пов'язують з початком розвитку обчислювальної техніки. На думку деяких авторів, ІТ є певні методичні системи (складові частини педагогічних технологій).

У деяких джерелах можна зустріти такі визначення ІТ навчання, які повністю не розкривають усі області їх застосування. Наприклад, ІТ навчання – процес підготовки і передачі інформації тому, хто навчається, за допомогою комп'ютера. Але у даному випадку правильніше було б визначати це як комп'ютерні, а не ІТ навчання, оскільки поняття ІТ набагато ширше.

У широкому сенсі під ІТ будемо розуміти сукупність методів і засобів збору, обробки і передачі первинної інформації (даних) для отримання нової якісної інформації про стан об'єкта, явища або процесу (інформаційного продукту).

Однією з основних особливостей ІТ у порівнянні з іншими областями науки і виробництва є те, що вони постійно змінюються. Це викликано бурхливим розвитком засобів сучасного зв'язку і комп'ютерної техніки. Сьогодні говорять не просто про ІТ, а про сучасні або новітні ІТ. За переконаннями деяких науковців, їх основу складають наступних п'ять технічних досягнень:

1. Поява нового середовища накопичення інформації.
2. Розвиток та широке охоплення засобів зв'язку дали можливість забезпечити доставку інформації в будь-яку точку земної кулі без значних обмежень у відстані та часі.
3. Можливість цифрової обробки інформації завдяки динамічному розвитку мікропроцесорної техніки.
4. Можливість автоматизованої обробки інформації за заданими алгоритмами за допомогою комп'ютера.
5. Виникнення і бурхливий розвиток мережі Інтернет [1].

Отже, новітніми ІТ навчання будемо називати сукупність засобів (у тому числі електронних) і способів їх функціонування, які можна використовувати для реалізації навчальної діяльності. Натомість, технології, які існували до появи мікропроцесорної обчислювальної техніки, будемо називати традиційними. Вони, як і раніше, успішно застосовуються при виробництві та в освіті і напевно чи будуть повністю витіснені новітніми ІТ.

Метою ІТ є виробництво інформації, яка повністю задовольняє інформаційні потреби людини. Дуже часто ці потреби пов'язані з прийняттям необхідних рішень у таких сферах, як спілкування, пізнання та практична діяльність.

Також варто зазначити, що однією з особливостей технології є те, що при застосуванні однієї і тієї ж технології до однієї і тієї ж сировини в результаті дасть «продукт» однієї і тієї ж якості. І навпаки: застосовуючи різні технології до однієї і тієї ж сировини, можна отримати різні «продукти».

Подібно до того, як в класичній технології можна виділити її складові (наприклад, проектування, матеріалознавство, виробничі процеси, тощо), так і ІТ можна розділити на складові частини: інформаційні процеси, теорія інформації, інформаційні системи, моделювання та формалізація, інформаційна безпека та інформаційне управління.

ІТ так само, як і будь-яка інша, повинна відповідати наступним вимогам: забезпечувати високий рівень поділу процесу обробки інформації на операції, дії, етапи; включати весь набір елементів, які є необхідними для рішення поставленого завдання; мати регулярний характер, тобто бути стандартизованими й уніфікованими, що дозволить більш ефективно здійснювати управління інформаційними процесами.

На сьогоднішній день існують різні підходи до класифікації ІТ. Наведемо декілька найбільш використовуваних класифікацій, наприклад: ІТ підтримки прийняття рішень; автоматизація офісу; ІТ управління; ІТ обробки даних; ІТ експертних систем [1].

Іншим видом класифікації є наступний: функціонально-орієнтовані ІТ (вирішують певні задачі); предметно-орієнтовані ІТ (їх основне призначення – вирішення конкретних завдань у межах певної предметної області); проблемно-орієнтовані ІТ (їх основне призначення – вирішення типових прикладних задач).

У рамках даної дисципліни буде використовуватися наступна класифікація, яка складена в залежності від форми подання оброблюваної інформації: технології роботи в глобальних мережах; технології обробки звукової інформації; технології обробки графічної інформації; технології обробки числової інформації; технології обробки текстової інформації; інформаційні технології комп'ютерного (технічного) зору.

Вибір даної класифікації пояснюється тим, що це - основні види ІТ, які вивчаються, починаючи зі школи.

В історії людського суспільства декілька разів відбувалися кардинальні зміни в галузі ІТ, які називаються інформаційними революціями.

Винахід писемності вважається першою інформаційною революцією. Писемність створила можливість поширення і накопичення знань для подальшої їх передачі майбутнім поколінням. Цивілізації, які опанували писемність, досягали вищого економічного і культурного рівнів, а також розвивалися швидше за інших. Прикладами таких цивілізацій можуть бути Древній Єгипет, Китай та інші. Перехід до алфавітного письма від піктографічного зробив писемність більш доступною та сприяв зміщенню центрів цивілізації в Європу (наприклад, в Грецію та Рим).

У середині XVI ст. відбулася друга інформаційна революція. Вона була пов'язана з винаходом друкарства. Стало можливим не просто зберігати інформацію, а й поширювати її, зробити масово-доступною. Грамотність стала масовим явищем. Все це прискорило розвиток науки і техніки, сприяло виникненню промислової революції.



У кінці XIX ст. відбулася третя інформаційна революція. Вона вже була обумовлена прогресом засобів зв'язку. Розробка радіо, телеграфу та телефону дозволили швидко передавати інформацію на великі відстані. Бурхливий розвиток природознавства також припадає на час третьої інформаційної революції.

Відносно недавно, у семидесятих роках XX ст., відбулася четверта інформаційна революція, яка пов'язана з розвитком електроніки та появою мікропроцесорної техніки. Одним з найбільш значущих винаходів того часу був персональний комп'ютер. Незабаром виникли комп'ютерні телекомунікації, а також через це радикально змінилися системи пошуку та зберігання інформації.

Завдяки значним змінам у суспільстві четверта інформаційна революція призвела до появи нового терміну – інформаційне суспільство. Ця назва вперше з'явилася в Японії. У момент виникнення даний термін визначав суспільство, в якому циркулює в достатку високоякісна інформація, а також є всі необхідні засоби для її використання, розподілу та зберігання. Інформація швидко і легко поширювалася у звичній формі за вимогами зацікавлених організацій та окремих людей.

Існує й інше визначення даного терміну. Під інформаційним (комп'ютеризованим) суспільством розуміється таке суспільство, усі сфери життя та діяльності якого включають використання комп'ютерної техніки або інших засобів інформатики, які є знаряддям інтелектуальної праці та відкривають широкий доступ до інформації, дозволяють з величезною швидкістю обробляти будь-яку інформацію, робити обчислення, моделювати поточні і прогнозувати майбутні події, процеси чи явища, автоматизувати навчання, управляти виробництвом, тощо.

У наш час жодна держава повноцінно не знаходиться в даній стадії, але ближче всіх до інформаційного суспільства підійшли США, Японія та ряд країн Західної Європи. Простежимо більш детально тенденції розвитку інформаційного суспільства, які існують зараз.

На сьогодні відсутній загальноприйнятий повноцінний та повномасштабний критерій оцінки інформаційного суспільства, проте є декілька його варіацій, наприклад, про етапи наближення до інформаційного суспільства варто судити за сукупною пропускну здатністю каналів зв'язку, розвиток яких відображає і об'єктивну потребу суспільства у всіх видах інформаційного обміну, і рівень його комп'ютеризації, а також інші прояви інформатизації. Відповідно до даного критерію, ранній етап інформатизації суспільства настає при досягненні сукупної пропускну здатності каналів зв'язку, яка забезпечує розгортання достатньо надійної міжміської телефонної мережі. Завершальний етап у цьому випадку настає за умови реалізації оперативного та надійного інформаційного контакту між членами суспільства за принципом «кожен з кожним». На завершальному етапі про-

пускна здатність каналів зв'язку повинна бути в мільйон разів більше, ніж на першому.

Варто також зазначити, що з кожним днем усе більше і більше людей в суспільстві починають працювати з інформацією. Інформація дозволяє людині відчувати себе частиною навколишнього світу, пізнавати його, виховувати дітей, спілкуватися з іншими людьми, вирішувати побутові проблеми, творчою працею, займатися діяльністю різного роду, тощо. Інформація допомагає в процесі організації роботи на підприємствах, при утворенні їх професійних спілок та товариств. Інформація є основою діяльності системи державного управління, органів виконавчої, законодавчої та судової влади.

Інформаційне суспільство – це суспільство, в якому більшість працюючих зайнято виробництвом, обробкою, реалізацією та зберіганням інформації, особливо вищої її форми – знань. Основною рушійною силою розвитку суспільства має стати інформаційне, а не матеріальне виробництво. Матеріальний продукт має включати в себе більше інформації, що призведе до збільшення частки інновацій, маркетингу, дизайну та його вартості.

Деякі науковці зазначають, що перехід до інформаційного суспільства США зможе завершити в до 2030 року, а Японія і більша частина країн Західної Європи до 2040 року. Зміни, які відбуваються при переході до інформаційного суспільства супроводжуються зміні пріоритетів в економіці з виробництва матеріальних благ (певних товарів) на надання послуг. Це в кінцевому результаті призведе до значного зниження витрат енергії, а також зниження видобутку і переробки сировини.

Завдяки інформатизації суспільства, друга половина ХХ століття супроводжувалася значним перетіканням людей зі сфери матеріального виробництва в інформаційну сферу. На сьогоднішній день промислові робітники в розвинених країнах становлять менше третини населення, проте в середині ХХ століття їх було більше 2/3 населення. Наприклад, до 1980 року 47% людей було зайнято в інформаційній сфері, 30% – у сфері обслуговування, 20% – у промисловості, а також 3% працюючих – у сільському господарстві.

Інформатизація змінила не тільки розподіл, але й характер праці в традиційних галузях промисловості. Розвиток робототехнічних систем та широке впровадження елементів мікропроцесорної техніки є основними причинами цього явища. Наприклад, у США в 1990 році в межах верстатобудівної галузі було задіяно приблизно 330 тисяч осіб, але до 2005 року залишилося лише близько 14 тисяч. Це пояснюється розвитком та масовим впровадженням засобів автоматизації виробництва, наприклад, промислових роботів та маніпуляторів.

Все це також призвело до виникнення ринку інформаційних продуктів і послуг, який постійно росте та розвивається. Цей ринок включає в себе

наступні сектори: послуг освіти; споживчої інформації (новини, розважальна інформація, тощо); професійної інформації (першоджерела, науково-технічна інформація, тощо); ділової інформації (фінансової, біржової, статистичної, комерційної); інші.

Однак у процесу інформатизації є також і недоліки, суть яких полягає в тому, що загальний потік інформації є настільки великим, що його неможливо обробити людині в прийнятний для неї час. Це так звана інформаційна криза. Це явище присутнє і в технічних розробках, і в наукових дослідженнях, і в суспільно-політичному житті.

Інформаційний криза найчастіше проявляється в наступному: виникають політичні, економічні та інші бар'єри, які перешкоджають розповсюдженню інформації (наприклад, через її секретність); виникає велика кількість надлишкової інформації («інформаційний шум»), яка ускладнює сприйняття корисної для споживача інформації; інформаційний потік є більшим за обмежені можливості людини в плані сприйняття та обробки інформації.

Частково вирішити проблему інформаційної кризи можна за рахунок застосування нових ІТ. Впровадження сучасних методів та засобів обробки, зберігання та передачі інформації значно знижують бар'єр доступу до неї і швидкість її пошуку. Звісно, одні лише ІТ не можуть повністю вирішити проблему, що має також економічний (деяка інформація коштує грошей), юридичний (інформація має власника), та ряд інших характеристик. Ця проблема являється комплексною, тому її вирішення потребує зусиль не тільки в межах кожної країни, а й у межах світової спільноти в цілому.

У наш час прийняття рішень з кожним днем стає все більш відповідальною справою, проте це неможливо зробити без повноти інформації. Прискорення процесу накопичення загального обсягу інформації відбувається з високою швидкістю. Загальний обсяг всієї інформації, яка вироблена людством, на початку ХХ століття подвоювався кожні 50 років. Для порівняння, до 1950 року подвоєння інформації відбувалося кожні 10 років, а до 1970 року – кожні 5 років. На даний момент цей процес не зупинився і ще більше прискорюється.

В основі інформаційної революції лежить вибуховий розвиток комунікаційних та ІТ. У цьому процесі можна чітко спостерігати також і зворотній зв'язок, тобто рух у напрямку до інформаційного суспільства сильно прискорює процеси розвитку зазначених технологій, роблячи їх широко використовуваними та затребуваними.

Зростання обсягів виробництва засобів обчислювальної техніки, яке почалося з середини ХХ століття, саме по собі не стало причиною переходу до інформаційного суспільства. Комп'ютери були доступними лише для порівняно невеликої кількості фахівців. Тож найважливішим етапом на шляху в інформаційне суспільство стало: вироблення єдиних правил поведінки в комп'ютерних мережах; пошук інформації в комп'ютерних

мережах; поява баз даних, доступ до яких через комп'ютерні мережі отримали мільйони людей по всьому світу; створення телекомунікаційної інфраструктури, яка включає в себе комп'ютерні мережі передачі даних.

Мабуть, найбільшу роль у цьому процесі зіграло створення всеоб'ємної комп'ютерної мережі Інтернет. Сьогодні вона являє собою найбільшу систему, яка дуже швидко зростає. В кожен момент часу в ній знаходиться понад 200 мільйонів чоловік одночасно.

Сьогодні у світі навіть спостерігається відмова від створення локальних корпоративних мереж на користь відкритих та стандартизованих систем з їх подальшою інтеграцією в мережу Інтернет. Винятком можуть бути мережі спеціального призначення, які висувають дуже високі вимоги до захисту корпоративної інформації.

Зараз розвиток комунікаційних та ІТ є постійним і неспинним. Провідні технології поступово стають більш універсальними. Це значить, що поступово з'являються потужні технології, які дозволять позбутися необхідності розробляти для кожного завдання окрему технологію.

Прикладом можуть виступати офісні програмні системи, за допомогою яких можна не тільки набирати текст, а й створювати спеціальні програми, наприклад, для нарахування заробітної плати за допомогою табличного процесора.

Широке використання мультимедіа також сприяє універсалізації ІТ. Сучасна мультимедійна система здатна об'єднати функції багатьох технічних засобів, таких як: телефон, автовідповідач, комп'ютер, телевізор, радіоприймач, факс, тощо. При цьому вона здатна забезпечувати також і доступ до мереж передачі даних.

Розвиток обчислювальної техніки призводить також до зменшення габаритних розмірів пристроїв зберігання інформації. Ці пристрої зараз можуть без проблем розміститися на долоні людини. Вони також мають деякі функції персонального комп'ютера, наприклад, дозволяють людині обзавестися власним універсальним довідником, обсяг інформації в якому можна порівняти з декількома реальними бібліотеками. Враховуючи той факт, що такі пристрої також мають підключення до мережі, вони також вміють передавати і оперативні дані, наприклад, поточний час, прогноз погоди, тощо.

Існує також термін «інформаційна культура». Найбільш часто даний термін вживається для позначення широти знань фахівця, проте визначення інформаційної культури в даний час остаточно не прийняте і найчастіше трактується по-різному.

Найчастіше інформаційну культуру фахівця характеризують з трьох сторін: знання, якими повинен володіти фахівець; якісна характеристика особистості; рівень інформатизації суспільства.

Людині, яка належить до інформаційного суспільства, необхідні такі навички та знання, які ефективно можна використовувати (для виявлення

потенціалу комп'ютерних технологій, подальшого просування техніки, науки, культури, тощо) та які повинні стати гарантом суверенізації особистості заради повної реалізації творчих ресурсів людини.

Комп'ютерну культуру можна умовно розділити на загальну та спеціальну. До загальної комп'ютерної культури можна віднести навички використання комп'ютерної техніки та створених для цього професійних прикладних програм.

До спеціальної комп'ютерної культури можна віднести ті знання, що забезпечують можливість фахівцеві працювати на стику своєї професії з обчислювальною технікою та інформатикою. Вона розвивається на вміннях використовувати комп'ютеризовані телекомунікаційні та ІТ у професійній діяльності фахівця, розумінні основних ідей інформатики, а також загальних навичок використання комп'ютерної техніки.

Майже для всіх учасників освітнього процесу комп'ютерна культура передбачає наявність знань, які забезпечують можливість студенту та викладачу користуватися основними поняттями інформатики та обчислювальної техніки під час навчання, а також володіння навичками вирішення завдань професійної діяльності з використанням комп'ютерів та спеціальних прикладних програм.

Розвиток комп'ютерної культури – особлива вимога до сучасного фахівця. Для вирішення цієї проблеми необхідно формувати нове мислення, що спирається на усвідомлення необхідності використання і застосування в будь-якому вигляді професійної діяльності сучасної комп'ютерної техніки, телекомунікаційних та ІТ, оволодіння якими можливо лише на базі комп'ютерної грамотності.

## **2. Особливості обробки багатовимірних статистичних даних. Види простору ознак.**

Обробка багатовимірних статистичних даних має певні особливості: застосування методів БСА вимагає творчого підходу до розв'язання аналітичних задач, оскільки оброблюються багатовимірні сукупності даних [2]. Для методів БСА характерна глибока формалізація та складна логіко-математична конструкція, їх практичне застосування потребує використання обчислювальної техніки (рис. 1).



**Рис. 1. Особливості застосування методів БСА**

Найбільш структурованою формою подання вихідних даних про багатовимірний об'єкт є матрична.

#### Види матриць:

1. Матриця значень аналітичних ознак ( $x_j$ ):

| Об'єкт | $X_1$    | $X_2$    | $X_3$    | $X_4$    | ... | $X_m$    |
|--------|----------|----------|----------|----------|-----|----------|
| $n_1$  | $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | ... | $x_{1m}$ |
| $n_2$  | $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | ... | $x_{2m}$ |
| $n_3$  | $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ | ... | $x_{3m}$ |
| ...    | ...      | ...      | ...      | ...      | ... | ...      |
| $n_n$  | $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ | ... | $x_{nm}$ |

2. Матриця теоретичних відстаней між об'єктами ( $n_j$ ):

| Об'єкт | $n_1$    | $n_2$    | $n_3$    | ... | $n_n$    |
|--------|----------|----------|----------|-----|----------|
| $n_1$  | $c_{11}$ | $c_{12}$ | $c_{13}$ | ... | $c_{1n}$ |
| $n_2$  | $c_{21}$ | $c_{22}$ | $c_{23}$ | ... | $c_{2n}$ |
| ...    | ...      | ...      | ...      | ... | ...      |
| $n_n$  | $c_{n1}$ | $c_{n2}$ | $c_{n3}$ | ... | $c_{nn}$ |

## Види простору ознак:

- з нульовою розмірністю (об'єкти не мають характеристик);
- одновимірний простір ознак (об'єкти відображаються значеннями однієї ознаки);
- багатовимірний простір (об'єкти подані значеннями двох і більше ознак).

Сучасні умови потребують від суб'єктів господарювання вміння аналізувати багатовимірні процеси та практично використовувати результати аналізу даних для прийняття ефективних управлінських рішень. Для того, щоб орган управління вибрав ту або іншу процедуру прийняття рішень (або механізм управління, тобто залежність своїх дій від цілей організації і дій суб'єктів управління – агентів), він має вміти передбачати поведінку агентів – їх реакцію на ті або інші управлінські дії (рис. 2).

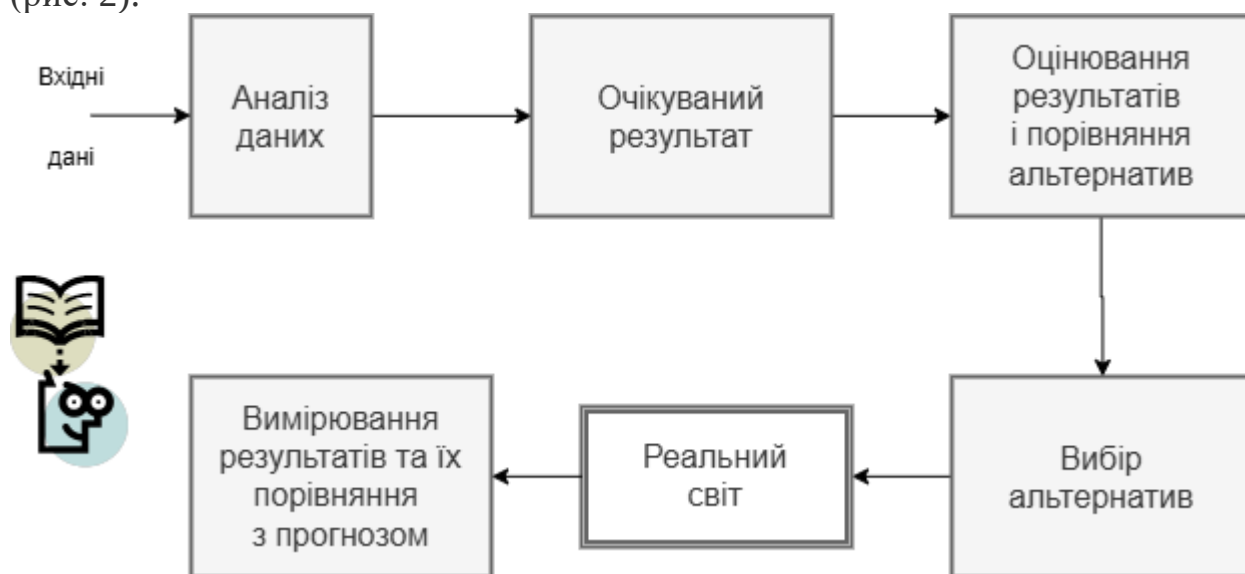


Рис. 2. Загальна схема процесу формування рішень

Експериментувати в житті, застосовуючи різні управлінські дії і вивчаючи реакцію підлеглих, неефективно та практично є неможливим. Тут на допомогу приходить моделювання – метод дослідження, що полягає в побудові й аналізі моделей – аналогів об'єктів дослідження. Маючи адекватну модель, можна з її допомогою проаналізувати реакції системи управління, а потім вибрати (на етапі синтезу) та використовувати на практиці той управлінський вплив, який призводить до необхідної реакції. Наявність моделей і механізмів управління є корисною як з точки зору органу управління (тому що дозволяє передбачити поведінку суб'єктів управління), так і самих суб'єктів (оскільки робить передбачуваною поведінку керівного органу). Тобто зниження невизначеності за рахунок використання механізмів управління

є однією з істотних властивостей будь-якої організації як соціального інституту.

Вивчення певного багатовимірного об'єкта будь-якої форми — це розкриття не тільки його якісних, але і кількісних закономірностей за допомогою моделей багатовимірних об'єктів. Під **моделлю** будемо розуміти образ реального об'єкта (процесу) в матеріальній або ідеальній формі (тобто описаний знаковими засобами будь-якою мовою), що відбиває істотні властивості змодельованого об'єкта (процесу) й управління в ході дослідження.

### **Проблема**

це ситуація, що характеризується розходженням між необхідним (бажаним) виходом та існуючим входом. Вона має два стани: існуючий і бажаний.

Проблеми розподіляють на:

- кількісні, для вирішення яких використовують способи маніпулювання числами;
- якісні проблеми, пов'язані з перерахуванням майбутніх чи погано визначених ресурсів і їх властивостей;
- слабоструктуровані проблеми, склад елементів, властивості та зв'язки яких відомі тільки частково.

### **Моделювання**

це універсальний спосіб вивчення процесів і явищ реального світу. Особливого значення воно набуває у процесі вивчення об'єктів, не доступних прямому спостереженню та дослідженню. До них, зокрема, відносять соціально-економічні явища та процеси. Моделювання завжди має цільову спрямованість. Цілі та методи його можуть бути різноманітними.

### **Економіко-математична модель**

це вираз формальної залежності у вигляді математичних співвідношень (функцій, систем рівнянь або нерівностей, операторів тощо) економічного процесу, проблеми або об'єкта, що досліджуються.

Надалі будемо говорити тільки про економіко-математичне моделювання методами багатовимірного аналізу, тобто про опис знаковими математичними засобами соціально-економічних систем.

Розрізняють вербальне, геометричне (предметне), фізичне й інформаційне моделювання:

- *вербальне моделювання* — моделювання на основі використання розмовної мови;

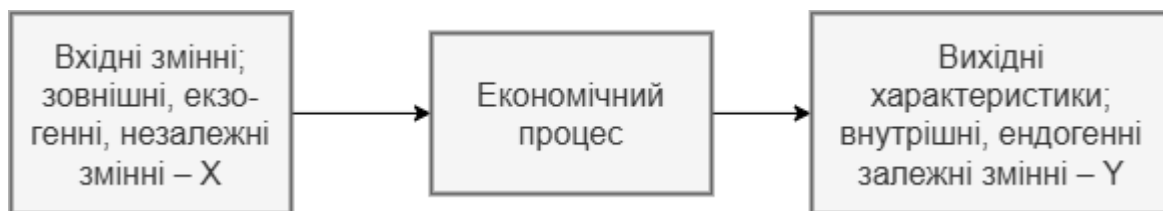


- *геометричне моделювання* здійснюється на макетах або об'єктних моделях. Ці моделі передають просторові форми об'єкта, пропорції тощо;
- *фізичне моделювання* застосовують для вивчення фізико-хімічних, технологічних, біологічних, генних процесів, що відбуваються в оригіналі. Таке моделювання називають аналоговим;
- *інформаційне моделювання* має фундаментальне значення в усіх галузях науки (схеми, графіки, креслення, формули, рівняння, нерівності).

Найважливіша роль серед методів інформаційного моделювання належить логіко-математичному моделюванню, тобто моделюванню за допомогою застосування математичного апарата багатовимірного статистичного аналізу.

Під час побудови економіко-математичної моделі реалізується метод моделювання за принципом «чорного ящика», коли досліднику невідомий механізм процесів, що протікають у системі, вивчити який можна за вхідними та вихідними характеристиками системи. Вхідні та вихідні характеристики системи часто ототожнюють з екзогенними й ендогенними змінними, або вживають терміни «незалежні» (факторні) змінні або ознаки і «залежні» (результативні) змінні або ознаки.

Графічно принцип «чорного ящика» зображено на рис. 3.



Вхідні змінні; зовнішні, екзо-генні, незалежні змінні – X

Вихідні характеристики; внутрішні, ендогенні залежні змінні – Y

**Рис. 3. Дослідження системи за принципом «чорного ящика»**

Досліднику необхідно виділити вхідні та вихідні характеристики та на підставі методів установити характер причинно-наслідкових зв'язків, що закладені в основу механізму функціонування соціально-економічної системи.

### **3. Етапи дослідження за допомогою багатовимірного статистичного аналізу:**

1. Постановка задачі: включає опис предметної області об'єктів, визначення обсягів виділених ресурсів (час, трудовитрати і т. д.).
2. Визначення набору методів багатовимірного статистичного аналізу та порядку їх використання.

3. Збирання вихідних даних, визначення способів збирання інформації, форм її подання.
4. Аналіз даних (перевірка однорідності вибірки, відповідність законам розподілу, виявлення грубих помилок і т. д.).
5. Уточнення математичної постановки задачі та визначення можливості застосування раніше відібраних методів (у разі необхідності набір методів змінюється).
6. Реалізація – проведення обчислень, реалізація за допомогою програмного забезпечення математичного інструментарію.
7. Оцінювання адекватності моделі, визначення несуперечності математичних результатів і економічних висновків.

Одним з головних принципів управління біотехнічними об'єктами є принцип системності. Система може бути подана нескінченним числом структурних і функціональних інваріантів, які відбивають взаємозв'язки між різними процесами, що протікають у цій системі (економічними, соціальними, екологічними, демографічними і т. д.). Опис системи здійснюється за допомогою її якісних і кількісних характеристик, іменованих параметрами. Параметри складають основу мов опису систем, а шляхом формалізації ототожнюються з незалежними змінними математичного опису процесу функціонування систем.

Чим повнішими й обґрунтованішими є принципи моделювання, тим вища точність моделі та ймовірність досягнення позитивних результатів від її застосування. У моделюванні найбільшого значення мають такі основні принципи.

**Принцип інтегратизму** полягає в тому, що окремі частини цілого характеризуються сукупністю трьох елементів:

1. Виникненням взаємодіючих систем – зв'язків між частинами цілого;
2. Утратою деяких властивостей частини зі входженням у ціле;
3. Появою нових властивостей у цілого, обумовлених властивостями складових.

Водночас обов'язковою є впорядкованість частин, детермінованість їх просторових і функціональних зв'язків, коли частина стає компонентом інтегрального цілого, внутрішньо об'єднаного. Цей принцип перетинається з відомим протилежним положенням У. Ешбі, який розглядає загальну теорію систем як загальну теорію спрощення.

**Принцип невизначеності** припускає, що «на краях» економічні процеси розпливчасті та невизначені. Протікаючи в часі, вони постійно змінюються, і якщо нам навіть вдасться встановити яку-небудь якість чи властивість процесу, то вона дійсна тільки в розглянутий момент часу в даній ситуації. Інакше кажучи, на мікрорівні економічні процеси необхідно розглядати з урахуванням випадкової зміни факторів.

**Принцип невизначеності** дозволяє також стверджувати, що існує рівень факторів, коли їхні малі відхилення не призводять до змін у стані системи.

Однак чим складніша модель системи, чим глибше ми намагаємося аналізувати її, тим більш невизначеним стає розв'язок задачі, а її результати віддаляються від практичного змісту.

**Принцип інваріантності** полягає в тому, що модель системи повинна бути інваріантною для будь-яких організаційних форм виробництва, а зміна умов не має змінювати сутності моделі.

**Принцип головних видів діяльності** полягає в тому, що в різних системах існують «схожі» види діяльності (управління, регулювання, розподіл і т. п.), які можна виділити як стандартні. Вони бувають незмінні на певному проміжку часу; їх можна описати деякими схожими моделями.

Для моделювання об'єктів і подання їх у вигляді систем необхідно враховувати **властивості складних систем** – такі, як:

- *емерджентність* як прояв у найбільш яскравій формі властивості цілісності системи. Тобто наявність у економічної системи таких властивостей, які не притаманні жодному з елементів системи, взятому окремо, поза системою. Емерджентність є результатом виникнення між елементами системи так званих синергічних зв'язків, що забезпечують збільшення загального ефекту до величини, більшої за суму ефектів елементів системи, які діють незалежно;
- *подільність* – цілісний об'єкт повинний бути зображений поділенням на елементи;
- *неможливість ізолювати процеси, що протікають в економічних системах*, і явища від навколишнього середовища, щоб спостерігати та досліджувати їх у чистому вигляді;
- *стійкість* – система має нормально функціонувати та бути нечутливою до неминучих сторонніх дестабілізаційних впливів;
- *розмаїтість* – кожному елементу системи властиве поведіння та стан, відмінні від поведінки та стану інших елементів і системи в цілому;
- *ідентифікованість* – кожен елемент системи може бути відділений від інших складових;
- *стабілізація* – система здійснює відновлення своїх елементів за рахунок їхнього регулювання;
- *спостережність* – усі без винятку входи та виходи системи або контрольовані дослідником, або принаймні дослідник може за ними спостерігати;
- *невизначеність* – дослідник одночасно не може фіксувати всі властивості та відношення елементів системи. Саме з метою їхнього виявлення він здійснює системне дослідження;
- *нетотожність відображення* – знакова система дослідника відмінна від знакової системи прояву властивостей об'єктів і їхніх

відношень. Втрата інформації визначає нетотожність системи досліджуваному об'єкту;

- *адаптація* – система зберігає стан рухливої рівноваги та стійкість до дестабілізаційних впливів, яким вона постійно піддається, шляхом перебудови внутрішньої структури та функцій окремих елементів;
- *масовий характер економічних явищ і процесів* – закономірності економічних процесів не виявляються на підставі незначної кількості спостережень;
- *динамічність економічних процесів* полягає в зміні параметрів і структури економічних систем під впливом середовища (зовнішніх факторів);
- *випадковість і невизначеність у розвитку економічних явищ* – закономірності економічних процесів виявляються на підставі випадкових величин і факторів, параметри розподілу яких можна встановити з деякою ймовірністю;
- *активна реакція на нові фактори, що з'являються* – здатність соціально-економічних систем до активних, не завжди передбачуваних дій залежно від відношення системи до цих факторів.

Контрольні запитання.

1. Що таке технологія?
2. Чим технологія відрізняється від інформаційної технології??
3. Як класифікуються інформаційні технології?
4. Які інформаційні революції відбулися у нашому суспільстві?
5. Яке суспільство називається інформаційним?
6. У чому проявляється інформаційна криза?
7. Особливості застосування методів БСА.
8. Етапи дослідження за допомогою багатовимірного статистичного аналізу.
9. Які основні принципи моделювання складних систем мають найбільшого значення?
10. Властивості складних систем.

## Лекція № 2

### Методи обробки багатовимірних даних у OLAP-системі

#### 1. Інформація, її уявлення та вимірювання

Чим є інформація? На дане питання немає однозначної відповіді. Термін «інформація» є контекстним, тобто його сенс залежить лише від контексту, в якому вони вживаються.

Будь-яка інформація має динамічний характер, тому що існує тільки в моменти взаємодії методів і даних. Весь інший час інформація перебуває в стані даних. Це призводить до того, що вона існує тільки в момент протікання інформаційного процесу.

У момент споживання одні й ті ж самі дані можуть надавати різну інформацію, яка залежить від ступеня адекватності взаємодіючих з ними методів. Наприклад, для студента, який не володіє певною мовою, написаний на цій мові текст дає тільки ту інформацію, яку можна отримати методом спостереження (наявність незнайомих символів, загальна кількість символів, тощо). Використання ж більш адекватних для даного випадку методів дасть іншу інформацію.

Дані є результатом реєстрації об'єктивно існуючих сигналів, які викликані змінами в полях або матеріальних тілах. Тому можна сказати, що дані є об'єктивними. На відміну від даних, методи є суб'єктивними. Методи також можна розділити на штучні та природні. В основі штучних методів лежать алгоритми – упорядковані послідовності команд – складені і підготовлені людьми (певними суб'єктами). В основі ж природних методів лежать біологічні властивості суб'єктів інформаційного процесу [1].

Отже, можна сказати, що інформація виникає і існує в момент взаємодії об'єктивних даних і суб'єктивних методів.

Характерною особливістю інформації є те, що на властивості інформації впливають властивості методів, які взаємодіють з даними в ході інформаційного процесу, і властивості даних, які складають її змістовну частину. Це відрізняє її від інших об'єктів природи і суспільства. Інформація має безліч властивостей. При цьому кожна наукова дисципліна розглядає ті властивості інформації, які є найбільш важливими в рамках відповідної області знань. Найбільш важливими (з точки зору інформатики) представляються наступні властивості: повнота, об'єктивність, адекватність, достовірність, доступність та актуальність інформації.

Повнота інформації. Повнота інформації часто характеризує її якість, а також визначає достатність наявних даних для створення на їх основі нових даних або для прийняття рішень. Чим більш повними є дані, тим більш широкий діапазон методів можна використовувати.

Об'єктивність і суб'єктивність інформації. Оскільки методи є суб'єктивними, можна сказати, що поняття об'єктивності є відносним. При-

йнято вважати більш об'єктивною саме ту інформацію, в яку методи вносять найменший елемент суб'єктивності.

**Адекватність інформації.** Під даним терміном найчастіше розуміється ступінь відповідності реальному об'єктивному стану речей. На основі неповних або недостовірних даних, а також через помилки в методології вимірювання (реєстрації) може утворюватися неадекватна інформація (навіть у випадку достовірних та повних даних).

**Достовірність інформації.** Дані виникають у момент вимірювання та реєстрації сигналів. Проте в більшості сигналів окрім корисної складової дуже часто можна знайти який рівень сторонніх сигналів, які не несуть корисної інформації в межах поставленої задачі. Такі елементи сигналів називаються інформаційним шумом. Тому основною задачею, яку необхідно вирішити для підвищення достовірності даних, є більш чітке виділення корисної складової серед інформаційного шуму. На сьогодні це можна зробити за допомогою великої кількості методів обробки даних (наприклад, фільтрації) або за допомогою використання новітніх методологій та засобів реєстрації сигналів (первинних перетворювачів).

**Доступність інформації.** Доступність являється мірою, яка характеризує можливість отримати певну інформацію. Доступність даних та доступність адекватних методів їх інтерпретації - основні фактори, які впливають на доступність інформації. Відсутність одного з цих факторів призводить до однакового результату, а саме: інформація виявляється недоступною.

**Актуальність інформації.** Актуальність характеризує ступінь відповідності інформації в поточний момент часу дійсності. Найчастіше інформаційні процеси являються розтягнутими в часі, тому достовірна та адекватна, але дуже застаріла інформація може приводити до прийняття помилкових рішень.

**Правильність відбору та формування інформації з метою адекватного відображення властивостей об'єкта** характеризує її репрезентативність. У даному випадку важливе значення мають: обґрунтованість відбору істотних зв'язків і ознак явища, яке розглядається; правильність концепції, на базі якої було сформульовано вихідне поняття.

До істотних погрішностей інформації нерідко призводить порушення її репрезентативності.

**Змістовність інформації** відображає семантичну ємність, рівну відношенню кількості семантичної інформації в повідомленні до обсягу оброблюваних даних.

**Точність інформації** найчастіше визначається ступенем близькості отриманої інформації до реального поточного стану об'єкта, явища, процесу, тощо.

**Стійкість** – це здатність інформації реагувати на зміни вихідних даних без порушення їх необхідної точності. Так само, як і репрезентативність,

стійкість інформації обумовлена правильністю вибору методики її відбору і формування.

Наведемо один з варіантів класифікації інформації:

1. За ступенем значущості інформація поділяється на: громадську; особисту; спеціальну.

Громадська інформація включає в себе звичайну, естетичну, суспільно-політичну та науково-популярну. Особиста інформація складається зі спадкової пам'яті конкретної людини, емоцій, почуттів, умінь, знань, досвіду, інтуїції, планів, прогнозів. Спеціальна - це наукова, технічна, виробнича, управлінська.

2. За способом сприйняття її можна розділити на: смакову; візуальну; нюхову; аудіальну; тактильну.

3. За способами (суб'єктам) обміну її можна також розділити на: генетичну; біологічну; технічну; соціальну.

4. За формою подання інформація поділяється на: звукову; графічну; числову; текстову.

У рамках іншої класифікації можна навести наступні категорії поділу: за характером носія інформації (паперовий, магнітний, інформація, яка зашифрована в молекулах ДНК, у довжинах світлових хвиль, тощо);

за характером джерел інформації (узагальнююча, первинна, вторинна, тощо);

за сферою застосування інформації (соціологічна, технічна, економічна, географічна, тощо).

Також розрізняють наступні види інформації в залежності від її типу носія: мовна (акустична); документальна; телекомунікаційна.

Мовна інформація виникає в ході процесу розмови, акустична – при роботі, наприклад, систем звуковідтворення та звукопідсилення. Звукові коливання в діапазоні частот (наближено) від 200 Гц до 5 кГц є носіями мовної інформації.

Документальна інформація може надаватися в графічному або буквено-цифровому вигляді на папері, а також в електронному вигляді на магнітних та інших носіях.

Телекомунікаційна інформація циркулює в технічних засобах обробки і зберігання інформації, а також в каналах зв'язку при її передачі. Носієм інформації при її обробці технічними засобами і передачі по провідним каналам зв'язку є електричний струм, а при передачі по радіо і оптичному каналах – електромагнітні хвилі.

Джерело інформації може виробляти безперервний сигнал, у цьому випадку інформація називається аналоговою (безперервною), або перервний (роздільний, дискретний) сигнал - інформація називається дискретною.

Форму безперервних залежностей, які швидко змінюються в часі, мають, наприклад, сигнали, що передаються по радіо і телебаченню, а також використовуються при магнітному записі, а також усі процеси, які

проходять у природі. На противагу цьому дискретні сигнали використовуються, наприклад, у телеграфії, комп'ютерній та обчислювальній техніці та мають переважно імпульсну форму.

Порівнюючи аналогову і дискретну форми подання інформації, неважко помітити, що при використанні аналогової форми для створення обчислювальної машини потрібно менше число пристроїв (кожна величина репрезентується одним, а не декількома сигналами), але ці пристрої будуть складнішими (вони повинні розрізняти значно більше число станів сигналу).

Інформація, що циркулює в суспільстві, вимагає спеціальних засобів і методів зберігання, використання та обробки. У процесі еволюції інформації сформувалися нові наукові дисципліни, наприклад, кібернетика, біоніка, робототехніка та інші, що мають на меті вивчення закономірностей інформаційних процесів.

Існує три підходи до вимірювання інформації:

1) Невимірювана інформація в побуті, при якому вважається, що будь-яка інформація є новою.

2) Технічний або об'ємний підхід – інформація представляється як повідомлення в формі знаків або сигналів, які оброблюються, перетворюються і зберігаються за допомогою технічних пристроїв.

В обчислювальній техніці застосовуються дві стандартні одиниці вимірювання інформації: біт і байт. Оскільки комп'ютер призначений для обробки великих обсягів інформації, то додатково використовують похідні одиниці - кілобайт (кб), мегабайт (Мб), гігабайт (Гб), тощо. У математиці префікс «кіло» означає тисячу, а префікс «мега» - мільйон. Але в обчислювальній техніці здійснюється прив'язка до прийнятої двійкової системи кодування.

У силу цього один кілобайт дорівнює не 1000 байтів, а  $2^{10}$ , тобто 1024 байта.

Аналогічно,  $1 \text{ Мб} = 2^{10} \text{ кб} = 1024 \text{ кб} = 2^{20} \text{ байтів} = 1\,048\,576 \text{ байтів}$ .

$1 \text{ Гб} = 2^{10} \text{ Мб} = 1024 \text{ Мб} = 2^{20} \text{ кб} = 2^{30} \text{ байтів} = 1\,073\,741\,824 \text{ байти}$ .

3) Імовірнісний підхід визначає, що інформація виступає у ролі знятої невизначеності.

Збільшення кількості інформації, тобто отримання нової, означає збільшення знання, що, в свою чергу, призводить до зменшення незнання або інформаційної невизначеності.

За одиницю кількості інформації найчастіше приймають вибір одного з двох рівномірних повідомлень, а саме: 1 або 0; так або ні. Ця одиниця також названа бітом.

Інформаційне моделювання і формалізація

Людство в своїй діяльності (науковій, освітній, технологічній, художній та інших) постійно створює і використовує моделі навколишнього світу. Суворі правила побудови моделей сформулювати неможливо, однак



людство накопичило багатий досвід моделювання різних об'єктів і процесів.

Термін «модель» в реальному житті має безліч значень.

Наприклад, модель – це:

- новий об'єкт, який відображає істотні з точки зору мети моделювання сторони досліджуваного об'єкта чи явища;
- відтворення предмета в зменшеному або збільшеному вигляді (макет);
- схема, зображення або опис якого-небудь явища або процесу в природі і суспільстві;
- спрощена подoba реального об'єкта;
- об'єкт-заступник, який в певних умовах може замінювати об'єкт-оригінал, відтворюючи основні його властивості і характеристики та має істотні переваги для дослідження або зручність використання;
- фізичний або інформаційний аналог об'єкта, функціонування якого за певними параметрами є подібним функціонуванню реального об'єкта;
- новий об'єкт (реальний, інформаційний або уявний), відмінний від вихідного, який володіє істотними для цілей моделювання властивостями і в рамках цих цілей повністю замінює вихідний об'єкт.

Все різноманіття моделей ділиться на три класи:

- матеріальні (натурні) моделі – зменшені або збільшені копії, які відтворюють зовнішній вигляд об'єкта, що моделюється, його структуру або поведінку (деякі реальні предмети - макети, муляжі, еталони, тощо);
- уявні моделі (геометрична точка, математичний маятник, ідеальний газ, нескінченність, тощо);
- інформаційні моделі – опис модельованого об'єкта на одній з мов кодування інформації (словесний опис, схеми, креслення, карти, малюнки, наукові формули, програми, тощо).

Наведемо класифікацію інформаційних моделей.

Інформаційна (абстрактна) модель – опис об'єкта на будь-якій мові. Абстрактність моделі проявляється в тому, що її компонентами є сигнали і знаки (вірніше, закладений в них сенс), а не фізичні тіла.

Дескриптивна (описова) модель – словесний опис об'єкта, виражений засобами тієї чи іншої мови.

Математична модель – сукупність записаних на мові математики співвідношень (формул, нерівностей, рівнянь, логічних співвідношень, тощо), що визначають характеристики стану об'єкта в залежності від його елементів, властивостей, параметрів, зовнішніх впливів; наближений опис об'єкта, виражений за допомогою математичної символіки.

Статичні моделі відображають об'єкт в якийсь момент часу без урахування змін, які відбуваються з ним, тобто відображають його таким, що знаходиться в стані спокою або рівноваги (відсутній параметр часу).

Динамічні моделі описують поведінку об'єкта в часі.

Детерміновані моделі відображають процеси, в яких відсутні випадкові величини.

Імовірнісні (стохастичні) моделі – опис об'єктів, поведінка яких визначається випадковими впливами (зовнішніми чи внутрішніми); опис імовірнісних процесів і подій, характер зміни яких у часі точно передбачити неможливо.

Імітаційна комп'ютерна модель – окрема програма, сукупність програм, програмний комплекс, що дозволяє за допомогою послідовності обчислень і графічного відображення їх результатів відтворювати (імітувати) процеси функціонування об'єкта, системи об'єктів за умови впливу на об'єкт різних факторів.

Імітаційна алгоритмічна модель – змістовний опис об'єкта в формі алгоритму, що відображає структуру і процеси функціонування об'єкта в часі, а також враховує вплив випадкових факторів.

Гносеологічна модель – опис об'єктивних законів природи.

Концептуальна модель описує виявлені причинно-наслідкові зв'язки і закономірності, притаманні досліджуваному об'єкту, і які є суттєвими в рамках певного дослідження.

Сенсуальні моделі – моделі почуттів, емоцій, або моделі, що впливають на почуття людини (музика, поезія, живопис, танець, тощо).

Модель-аналог – аналог об'єкта, який веде себе як реальний об'єкт, але не виглядає як такий.

Моделювання – це:

- дослідження об'єктів пізнання за їх моделями;
- заміна реального об'єкта його подібним аналогом;
- побудова моделей реально існуючих об'єктів (предметів, явищ, процесів, тощо).

Потреба в моделюванні виникає в таких сферах людської діяльності як: пізнання, практична діяльність, спілкування. Аспекти моделювання характеризуються властивостями:

- поведінка – зміна зовнішнього вигляду і структури з плином часу;
- структура – перелік елементів і зазначення відносин між ними;
- зовнішній вигляд - набір ознак.

Визначимо етапи моделювання:

1. Постановка цілей моделювання.
2. Аналіз об'єкта моделювання, виділення всіх його відомих властивостей.
3. Аналіз виділених властивостей і визначення істотних з них.
4. Вибір форми представлення моделі.
5. Формалізація.
6. Аналіз отриманої моделі на суперечливість.

7. Аналіз адекватності отриманої моделі об'єкту і мети моделювання.

Формалізація – це приведення істотних властивостей і ознак об'єкта моделювання до обраної форми.

Щоб побудувати модель, необхідно надати об'єкту форму. Суть формалізації полягає в принциповій можливості поділу об'єкта і його позначення. Для того, щоб позначити об'єкт, потрібно ввести деякий набір знаків.

Знак – це елемент кінцевої множини відмінних один від одного елементів. Слід звернути увагу на те, що поняття знака є одним з базисних понять науки. Точного визначення дати неможливо. Тому варто обмежитися зазначенням основних рис знака:

- багатозначність відповідності «знак - денотат» (об'єкт позначення);
- здатність знака виступати в якості замітника об'єкта (денотата);
- нетотожність знака і денотата – знак ніколи не може повністю замінити денотат.

Мова – знакова система, яка використовується з метою пізнання і комунікації. Слід розглянути характеристики мови і вказати, що мови можуть бути природними і штучними. Правила штучної мови є строго і однозначно певними, тому така мова називається формалізованою.

Процес формалізації текстової інформації (подання інформації у формі графа, креслення, схеми, тощо) здійснюється з метою її однозначного розуміння, полегшення і прискорення її обробки. Формалізувати можна і оформлення тексту. Цей процес полягає в використанні бланків, формулярів, шаблонів заздалегідь визначеної або затвердженої форми (стандарту).

Таблиці – форма подання інформації в зручному для аналізу і обробки вигляді. Таблиці бувають типу «об'єкт-об'єкт», «об'єкт-властивість», «об'єкт-властивість-об'єкт». Таблиця характеризується назвою, кількістю стовпців і їх назвами, кількістю рядків і їх назвами, вмістом комірок.

Граф – сукупність точок, з'єднаних між собою лініями. Ці точки називаються вершинами графа. Лінії, що з'єднують вершини, називаються дугами, якщо заданий напрямок від однієї вершини до іншої, або ребрами, якщо спрямованість двостороння.

## **2. Методи обробки багатовимірних даних у OLAP-системі**

Системи аналізу даних, побудовані на основі OLAP (Online Analytical Processing – оперативний аналіз даних), дозволяють виконувати швидкий і ефективний аналіз над великими об'ємами даних [2]. Дані зберігаються в багатовимірному вигляді, що найбільш близько відображає природний стан реальних даних. Засоби OLAP-аналізу можуть добувати дані безпосередньо з реляційних систем і з багатовимірних СУБД.

Основними методами аналітичної обробки даних в KIC є: підтримка прийняття рішень через аналіз даних – OLAP; Knowledge Discovery in Databases – виявлення знань у базах даних; Data Mining – добування даних.

СБЕАД (система багатовимірного експрес-аналізу даних), на відміну від статичних систем, дозволяє користувачу динамічно оперувати даними, швидко створювати нові звіти і переглядати їх результати.

Основна мета, яку необхідно досягнути при розробці СБЕАД – швидкість опрацювання даних. Це дозволяє використовувати їх в процесі інтерактивної роботи аналітика з інформацією. Тому, необхідно використовувати OLAP-системи. Першим кроком при розробці архітектури СБЕАД є аналіз запитів і вимог, пропонованих користувачами до СБЕАД. Було проведено декілька консультацій зі співробітниками підприємства, які планують використати СБЕАД і з'ясовано перелік вимог до системи. Було визначено первісний перелік питань, відповіді на які необхідно буде одержати.

Найбільше значення мають не самі формулювання питань, а перелік фактів, подій, об'єктів, які будуть потрібні для відповідей на такі питання, і які необхідно зберігати в сховищі даних. OLAP-клієнт може бути реалізований декількома методами. По-перше, можна використати ActiveX-елемент, що входить до складу Microsoft Office і забезпечує можливість аналізу багатовимірних даних та відображення різних звітів. По-друге, мовою високого рівня може бути написаний клієнтський додаток, що виконує віддалений доступ до сховища даних, реалізуючи при цьому будь-які специфічні потреби користувачів. Потретє, можна використати інтерфейс ADO MD (Active Data Objects MultiDimensional), завдяки якому можна створювати Веб-орієнтовані додатки.

Для застосування OLAP-технології при реалізації СБЕАД необхідно проаналізувати поняття сховища даних, основні елементи OLAP та визначити архітектуру OLAP-додатків. Також доцільно провести аналіз реалізації СБЕАД в розподілених середовищах та запропонувати вирішення проблеми проектування вискоєфективних паралельних архітектур серверів БД (симетричні багатопроцесорні системи (SMP) – найбільш часто використовується форма сильно зв'язаних багатопроцесорних систем, тобто систем, що розділяють єдину оперативну пам'ять і найбільш часто – дискову підсистему; слабозв'язані багатопроцесорні системи – сукупність самостійних комп'ютерів, об'єднаних у єдину систему швидкодіючою мережею й, можливо, що мають загальну дискову підсистему, як, наприклад, кластерні інсталяції; системи з масовим паралелізмом (MPP) – системи із сотнями й навіть тисячами процесорів, деталі їхньої реалізації можуть значно розрізнятися).

Існують чотири групи вимог, які визначають з технічної точки зору споживачів якості сучасної СУБД: масштабування; продуктивність; можливість змішаного завантаження різними типами завдань; забезпечення постійного доступу даних [2]. Для досягнення максимальної продуктивності СУБД при змішаному завантаженні необхідно використати оптимізатор запитів, керування ресурсами й паралельною обробкою

запитів. Процес проектування багатовимірної БД складається з декількох етапів: визначення запитів та вимог до системи; визначення структури БД (виміри й факти, їхні взаємозв'язки й рівні агрегації збережених даних); визначення реляційної схеми («сніжинка» – дозволяє заощадити дисковий простір, збільшується швидкість перегляду вимірів); реалізація багатовимірної БД в MS OLAP Services. Останній етап повинен бути виконаний у такій послідовності: створення БД – New Database; створення джерела даних – Data Sources; створення кубів – Cube Wizard; редагування створеної схеми куба (за потреби) – Cube Editor; проектування фізичного сховища – Design Storage; визначення ступені «агрегації» куба – Storage Design Wizard; завантаження даних в куб –Process Now. Після цього їх можна переглянути засобом Cube Browser. Проблема впровадження СБЕАД у середовище WWW стоїть досить гостро з кількох причин. По-перше, використання в якості програми-клієнта звичайного Веб-браузера дозволяє знизити вимоги до системних ресурсів клієнтського робочого місця. По-друге, якщо буде потреба внесення будь-яких змін в інтерфейс або в програмну частину, необхідно змінити програму лише на Веб-сервері. Після цього всі користувачі відразу одержать доступ до оновленої версії СБЕАД. По-третє, використання мережі Internet дозволить мати доступ до СБЕАД, хоча це, у свою чергу, має на увазі необхідність створення надійних засобів забезпечення безпеки. По-четверте, WWW-технологія створювалася для забезпечення багатокористувацької роботи, що також є однією з основних вимог до СБЕАД. Реалізація Веб-клієнта для доступу до СБЕАД можлива, наприклад, із застосуванням інтерфейсу ADO MD, що використовує технологію Active. ASP-програма, написана мовою VB Script, дозволяє будь-якому клієнтові мережі Internet з'єднуватися з OLAP сервером, що перебуває за тисячі кілометрів від нього, і за допомогою інтерфейсу ADO через спеціальний OLE DB Provider для OLAP переглядати таблиці з аналітичними даними. Дана програма під'єднується до OLAP-сервера й по заданому MDX-запиті показує результати в HTML-таблиці. Для виконання програми з HTML-браузера, треба попередньо скопіювати її в каталог WEB-сервера. Потім потрібно за допомогою OLAP-менеджера встановити права читання інформаційного куба для всіх користувачів, що підключаються, у режимі читання. Цю програму можна виконати із браузера з будь-якого комп'ютера локальної мережі. Наступний варіант реалізації www-клієнта для OLAP-сервера простіший, ніж попередній. У веб-сторінку встановлюється ActiveX-компонент Pivot Table, що на основі інтерфейсу OLE DB for OLAP здійснює взаємодію з OLAP-сервером. Для реалізації такого клієнта необхідно мати на серверному комп'ютері MS Office із установленим Microsoft Front Page. Однак це призводить до необхідності установки даного ActiveXкомпонента на кожному із клієнтських місць. Це є обмеженням, але тільки лише у випадку, якщо надається доступ клієнтів через мережу Internet. У випадку роботи в мережі

Intranet, що є більше ймовірним для СППР, ця проблема знімається, тому що не є складно встановити ActiveX-компонент на декількох клієнтських комп'ютерах, на яких користувачі будуть робити аналіз даних.

Контрольні запитання.

1. Що таке інформація?
2. Які є властивості у інформації?
3. Як класифікується інформація?
4. Які існують підходи до вимірювання інформації?
5. Що таке модель?
6. Як класифікуються моделі?
7. Що таке формалізація?
8. Основні методи аналітичної обробки даних в КІС.
9. Чотири групи вимог до якості сучасної СУБД.
10. Основні етапи проектування багатовимірної БД.
11. Які існують варіанти реалізації www-клієнта для OLAP-сервера?

## Лекція № 3

### Класифікація даних з використанням детермінованих моделей

#### 1. Інформаційні процеси та інформаційні системи

Наступний важливий компонент інформаційних технологій, який недостатньо добре освітлений, стосується інформаційних процесів. Інформація не існує сама по собі, вона проявляється в інформаційних процесах.

У найбільш загальному вигляді інформаційний процес визначається як сукупність послідовних дій (операцій), що проводяться над інформацією (у вигляді даних, відомостей, фактів, ідей, гіпотез, теорій, тощо) для отримання будь-якого результату або досягнення певної мети.

Інформаційні процеси можуть бути цілеспрямованими або стихійними, організованими або хаотичними, детермінованими або імовірнісними. Слід звернути увагу на те, що інформаційний процес завжди протікає в будь-якій інформаційній системі - біологічній, соціальній, технічній, соціо-технічній.

Залежно від того, якого роду інформація є предметом інформаційного процесу і хто є його суб'єктом (технічний пристрій, людина, колектив, суспільство в цілому), можна говорити про глобальні інформаційні процеси (мікропроцеси) і локальні інформаційні процеси (мікропроцеси).

Найбільш загальними інформаційними процесами є три процеси: збір, перетворення, використання інформації. Кожен з цих процесів розпадається, в свою чергу, на ряд менших процесів, причому деякі з останніх можуть входити в кожен з виділених узагальнених процесів.

Так, збір інформації складається з процесів відбору і пошуку. У той же час пошук інформації здійснюється в результаті використання конкретних методів пошуку й виконання процедур визначення мети.

Методи пошуку бувають «ручні» або автоматизовані. Вони включають у себе такі процедури, як формування пошукового образу (в явному або неявному вигляді), перегляд інформації, що надходить, з метою порівняння її з пошуковим запитом.

Відбір інформації проводиться на основі її аналізу і оцінки її властивостей відповідно до обраного критерію оцінки. Відібрана інформація найчастіше зберігається.

Зберігання інформації – це розповсюдження її в часі. Зберігання інформації неможливе без виконання процесів кодування, формалізації, структуризації, розміщення, що відносяться до загального процесу перетворення інформації.

У свою чергу кодування, формалізацію, структурування цілком обґрунтовано можна віднести до процесів обробки інформації. Поряд з перерахованими вище, до процесів обробки інформації відносяться також ін-

формаційне моделювання, обчислення за формулами (чисельні розрахунки), узагальнення, систематизація, класифікація, схематизація, тощо.

Обробка інформації становить основу процесу перетворення інформації.

Інформація може бути передана (поширена в просторі) для її подальшого використання, обробки або зберігання. Процес передачі інформації включає в себе процеси кодування, сприйняття, розшифровки, тощо.

Найважливішим процесом використання інформації суб'єктом є процес підготовки і прийняття рішень. Поряд з цим часто використання інформації зводиться до процесів формування документованої інформації з метою підготовки інформаційного або керуючого впливу.

У реальній практиці широко використовуються процедури, що входять у процес захисту інформації.

Захист інформації – важливий компонент процесів зберігання, обробки, передачі інформації в системах будь-якого типу, особливо в соціальних і технічних. До неї відносяться: розробка коду (шифру), кодування (шифрування), порівняння, аналіз, пароллювання, тощо [4].

Після того, як процес використання інформації завершено, наприклад, рішення прийнято і суб'єкт приступив до його реалізації, як правило, виникає нова задача і необхідна нова інформація або уточнення вже існуючої. Це призводить до того, що суб'єкт знову звертається до процедури збору інформації. Тому, говорячи про інформаційні процеси, слід підкреслювати не тільки їх взаємозв'язок, а й циклічність.

Також варто звернути увагу на такі способи обробки інформації, як: аналіз, синтез, порівняння, групування, опорні пункти, мнемонічний план, структурування, систематизація, схематизація, аналогія, асоціація, тощо.

Аналіз – метод наукового дослідження, шляхом розгляду окремих сторін, властивостей, складових частин чого-небудь.

Синтез – метод дослідження будь-якого явища в його єдності і взаємного зв'язку частин, узагальнення, зведення в єдине ціле даних, здобутих аналізом.

Порівняння – процес зіставлення для встановлення подібності або відмінності.

Групування – розбиття матеріалу на групи по будь-яким категоріям (по змісту, асоціаціям, тощо).

Опорні пункти – виділення будь-якого короткого пункту, що служить опорою ширшого змісту (тези, заголовки, питання, образи, приклади, цифрові дані, порівняння, імена, епітети, тощо). Сам набір опорних пунктів є інструментом або знаряддям запам'ятовування або відтворення іншого порядку, де закодований весь матеріал.

Мнемонічний план – сукупність опорних пунктів. У ньому можуть відбиватися і зовнішні зв'язки, і внутрішні, що характеризують відносини



різних груп матеріалу і смислові зв'язки з наявними в тексті даними, зв'язки з особистим досвідом, знаннями і цінностями суб'єкта.

Класифікація – розподіл будь-яких об'єктів, явищ, понять по класах, групах, розрядах на основі певних загальних ознак. При побудові класифікації можна виділяти: склад, структуру, кількісні характеристики, умови і причини виникнення, етапи розвитку, тощо.

Структурування – процес встановлення взаємного розташування частин, що складають ціле, визначення внутрішньої будови матеріалу.

Систематизація – встановлення певного порядку в розташуванні частин цілого і зв'язків між ними.

Схематизація – зображення або опис чого-небудь в основних рисах або спрощене уявлення інформації.

Аналогія – встановлення подібності, подоби в певних відносинах предметів, явищ, понять, які в цілому різні.

Асоціація – встановлення зв'язків за подібністю, суміжністю, протилежністю, тощо.

З інформаційними процесами дуже тісно пов'язане таке поняття, як інформаційна система.

Інформаційна система – взаємозв'язана сукупність засобів, методів і персоналу, використовуваних для зберігання, обробки та видачі інформації в інтересах досягнення поставленої мети.

Інформаційна система є інструментарієм інформаційних технологій та немислима без персоналу, що взаємодіє з комп'ютером і телекомунікаціями.

По виду підтримуваних інформаційних технологій, за класами вирішуваних завдань і по областях застосування можна виділити наступні інформаційні системи:

- системи, що підтримують бази даних;
- системи обробки даних;
- обчислювальні системи;
- системи автоматизованого проектування;
- документографічні системи, що підтримують повнотекстові документальні архіви;
- геоінформаційні системи;
- видавничі системи;
- системи інформаційного обслуговування (системи науково-технічної інформації, інформаційно-пошукові системи, тощо);
- системи діагностики.

Властивості інформаційних систем:

- при побудові інформаційних систем необхідно використовувати системний підхід;
- будь-яка інформаційна система може бути піддана аналізу, побудована і керована на основі загальних принципів побудови систем;

- інформаційну систему варто сприймати як людино-комп'ютерну систему обробки інформації;
- інформаційна система є динамічною і розвивається;
- вихідною продукцією інформаційних систем є інформація, на основі якої приймаються рішення.

Загальну структуру інформаційних систем можна розглядати, як сукупність підсистем, незалежно від сфери застосування. Структура будь-якої інформаційної системи може бути представлена сукупністю підсистем: інформаційної, програмної, організаційної, технічної, математичної, правової, тощо.

Також слід зазначити, що інформаційні системи можуть бути «ручними», автоматичними і автоматизованими. «Ручні» інформаційні системи характеризуються відсутністю сучасних технічних засобів перетворення інформації та виконанням всіх операцій людиною. Автоматичні інформаційні системи виконують всі операції по перетворенню інформації без участі людини. Автоматизовані інформаційні системи передбачають участь в процесі обробки інформації і людини, і технічних засобів, причому основним засобом є комп'ютер.

Технічна база інформаційної технології

Розвиток електронних обчислювальних машин можна умовно розбити на кілька етапів (поколінь електронних обчислювальних машин або ЕОМ), які мають свої характерні особливості.

Перший етап (ЕОМ першого покоління) - до кінця 50-х років XX століття.

Точкою відліку ери ЕОМ вважають 1946 рік, коли був створений перший електронний комп'ютер ENIAC (Electronic Numerical Integrator and Computer). Обчислювальні машини цього покоління будувалися на електронних лампах, які споживають величезну кількість електроенергії і виділяють багато тепла.

Числа в ЕОМ вводилися за допомогою перфокарт і набору перемикачів, а програма задавалася з'єднанням гнізд на спеціально набраних платах. Продуктивність такої гігантської ЕОМ була нижче за сучасний калькулятор. Широкому використанню таких ЕОМ, крім високої вартості, перешкоджали також низька надійність, обмеженість їх ресурсів і надзвичайно трудомісткий процес підготовки, введення і налагодження програм, написаних на мові машинних кодів. Основними їх користувачами були вчені, що вирішували найбільш актуальні науково-технічні завдання, пов'язані з розвитком реактивної авіації, ракетобудування, тощо.

Другий етап (ЕОМ другого покоління) - до середини 60-х років XX століття.

Розвиток електроніки призвів до винаходу в 1948 р. нового напівпровідникового пристрою - транзистора, який замінив лампи. Творцями транзистора є американські фізики У. Шоклі, В. Браттейн і Дж. Бардін за

цей винахід були удостоєні Нобелівської премії. Поява ЕОМ, побудованих на транзисторах, призвела до зменшення їх габаритів, маси, енергоспоживання і вартості, а також до збільшення їх надійності та продуктивності. Однією з перших транзисторних ЕОМ була створена в 1955 році бортова ЕОМ для міжконтинентальної балістичної ракети ATLAS.

Зі структурної точки зору ЕОМ другого покоління характеризуються розширеними можливостями по введенню-висновку, збільшеним об'ємом запам'ятовуваних пристроїв, розвиненими системами програмування.

У цей період з'явилися так звані алгоритмічні мови високого рівня, кошти яких допускають опис всієї необхідної послідовності обчислювальних дій в наочному вигляді, який легко сприймається.

З'явився широкий набір бібліотечних програм для вирішення різноманітних математичних задач. Були створені моніторні системи, керівники режимами трансляції та виконання програм. З моніторних систем надалі вирости сучасні операційні системи (ОС).

Перші ОС просто автоматизували роботу оператора ЕОМ, пов'язану з виконанням завдання користувача: введення в ЕОМ тексту програми, виклику потрібного транслятора, виклик необхідних бібліотечних програм, тощо. Тепер же разом з програмою і даними в ЕОМ вводиться ще й інструкція, де перераховуються етапи обробки і приводиться ряд відомостей про програму та її автора. Потім в ЕОМ стали вводити відразу по кілька завдань користувача (пакет завдань), ОС стали розподіляти ресурси ЕОМ між цими завданнями - з'явився мультипрограмний режим обробки.

Третій етап (ЕОМ третього покоління) – до початку 70-х років ХХ століття.

Елементною базою в ЕОМ третього покоління є інтегральні схеми. Створення технології виробництва інтегральних схем, що складаються з десятків електронних елементів, утворених у прямокутній пластині кремнію з довжиною сторони не більше 1 см, дозволило збільшити швидкодію і надійність ЕОМ на їх основі, а також зменшити габарити, споживану потужність і вартість ЕОМ.

Машини третього покоління – це сімейство машин з єдиною архітектурою, тобто програмно-сумісних. Вони мають розвинені операційні системи, володіють можливостями мультипрограмування. Багато завдань управління пам'яттю, пристроями і ресурсами стала брати на себе ОС або ж безпосередньо сама машина.

Швидкодія машин всередині сімейства змінюється від декількох десятків тисяч до мільйонів операцій в секунду. Ємність оперативної пам'яті досягає декількох сотень тисяч слів.

У цей період широкого поширення набули міні-ЕОМ. Простота обслуговування міні-ЕОМ, їх порівняно низька ціна і малі габарити дозволяли забезпечити цими машинами невеликі колективи дослідників, розробників-експериментаторів, тощо.

У період машин третього покоління сталося велике зрушення в області застосування ЕОМ. Якщо раніше ЕОМ використовувалися в основному для науково-технічних розрахунків, то в 60-70 роки все більше місця стала займати обробка символічної інформації.

Четвертий етап (ЕОМ четвертого покоління) – по теперішній час.

Цей етап умовно ділять на два періоди: перший – до кінця 70-х років і другий – з початку 80-х років по теперішній час.

У перший період успіхи в розвитку електроніки привели до створення великих інтегральних схем (ВІС), де в одному кристалі розміщувалося кілька десятків тисяч електронних елементів. Це дозволило розробити більш дешеві ЕОМ, які мали велику пам'ять та менший цикл виконання команд: вартість байта пам'яті та однієї машинної операції різко знизився. Так як витрати на програмування майже не скорочувалися, то на перший план вийшло завдання економії людських, а не машинних ресурсів.

Розроблялися нові ОС, що дозволяли програмістам налагоджувати свої програми прямо за дисплеєм ЕОМ, що прискорювало розробку програм.

У 1971 р. був виготовлений перший мікропроцесор – велика інтегральна схема, в якій повністю розміщувався процесор ЕОМ простої архітектури. Стала реальною можливість розміщення в одній ВІС майже усіх електронних пристроїв нескладної по архітектурі ЕОМ, тобто можливість серійного випуску простих ЕОМ малої вартості. З'явилися дешеві мікрокалькулятори і мікроконтролери - керуючі пристрої, побудовані на одній або декількох ВІС, що містять процесор, пам'ять і системи зв'язку з датчиками та виконавчими органами в об'єкті управління.

У іншому періоді поліпшення технології ВІС дозволяло виготовляти дешеві електронні схеми, що містять сотні тисяч елементів у кристалі - надвеликі інтегральні схеми (НВІС).

З'явилася можливість створити настільний прилад з габаритами телевізора, в якому розміщувалися мікро-ЕОМ, клавіатура, а також схеми сполучення з малогабаритним друкуючим пристроєм, вимірювальною апаратурою, іншими ЕОМ, тощо. Завдяки ОС, що забезпечує простоту спілкування з цією ЕОМ, великій бібліотеці прикладних програм з різних галузей людської діяльності, а також малої вартості такий персональний комп'ютер стає необхідною приналежністю будь-якого фахівця.

З точки зору структури машини це покоління є багатопроцесорними та багатомашинними комплексами, що працюють на загальній пам'яті та загальному полі зовнішніх пристроїв. Для цього періоду характерні широкі застосування систем управління базами даних (СУБД), комп'ютерних мереж, систем розподіленої обробки даних.

Класифікація засобів обробки інформації

Існують різні системи класифікації електронних засобів обробки інформації: по архітектурі, по продуктивності, за умовами експлуатації, за кількістю процесорів, за споживчими властивостями, тощо. Один з най-

більш ранніх методів класифікації - класифікація по продуктивності та характеру використання комп'ютерів. Відповідно до цієї класифікації комп'ютерні засоби обробки можна умовно розділити на наступні класи:

- суперкомп'ютери;
- мейнфрейми;
- мікрокомп'ютери.

Мікрокомп'ютери. Спочатку визначною ознакою мікрокомп'ютера була наявність у ньому мікропроцесора, тобто центрального процесора, який був виконаний у вигляді однієї мікросхеми. Зараз мікропроцесори використовують у всіх без винятку класах ЕОМ, а до мікрокомп'ютерів відносять більш компактні в порівнянні з мейнфреймами ЕОМ, що мають продуктивність до сотень мільйонів команд в секунду.

Сучасні моделі мікрокомп'ютерів володіють декількома мікропроцесорами. Продуктивність комп'ютера визначається не тільки характеристиками застосовуваного мікропроцесора, але й ємністю оперативної пам'яті, типами периферійних пристроїв, якістю конструктивних рішень, тощо.

Мікрокомп'ютери представляють собою інструменти для вирішення різноманітних складних завдань. Їх мікропроцесори з кожним роком збільшують потужність, а периферійні пристрої – ефективність.

Персональні комп'ютери (ПК) – це мікрокомп'ютери універсального призначення, розраховані на одного користувача і керовані однією людиною. У клас персональних комп'ютерів входять різні обчислювальні машини – від дешевих домашніх та ігрових з невеликою оперативною пам'яттю до надскладних машин з потужним процесором, накопичувачем місткістю в десятки гігабайт, з кольоровими графічними пристроями високої роздільної здатності, засобами мультимедіа та іншими додатковими пристроями.

ПК можна класифікувати і за конструктивними особливостями. Вони підрозділяються на стаціонарні (настільні) і переносні. Переносні, в свою чергу, діляться на портативні (Laptop), блокноти (Notebook), кишенькові (Palmtop).

Мейнфрейми. Призначені для вирішення широкого класу науково-технічних завдань і є складними і дорогими машинами. Їх доцільно застосовувати в більших системах при наявності не менше 200-300 робочих місць. Кілька мейнфреймів можуть працювати спільно під керуванням однієї ОС над виконанням єдиного завдання.

Суперкомп'ютери. Це дуже потужні комп'ютери з продуктивністю понад 100 Мфлопс (1 Мфлопс – це сто мільйонів операцій в секунду). Вони називаються надшвидкодіючими. Створити такі високопродуктивні ЕОМ за сучасною технологією на одному мікропроцесорі не представляється можливим через обмеження, обумовлені кінцевим значенням швидкості поширення електромагнітних хвиль, так як час поширення сигналу на відстань кількох міліметрів (лінійний розмір ребра мікропроцесора) при

швидкодії 100 млрд оп./с стає порівняним з часом виконання однієї операції. Тому супер ЕОМ створюються у вигляді високопаралельних багато-процесорних обчислювальних систем.

Персональний комп'ютер - універсальна технічна система. Його конфігурацію можна гнучко змінювати в міру необхідності. Проте, існує поняття базової конфігурації, яку вважають типовою. У даний час в базовій конфігурації розглядають чотири пристрої:

- системний блок;
- монітор;
- клавіатура;
- миша.

Системний блок являє собою основний вузол, усередині якого встановлені найбільш важливі компоненти. Пристрої, що знаходяться всередині системного блоку, називають внутрішніми, а пристрої, що підключаються до нього зовні, - зовнішніми. Зовнішні додаткові пристрої, призначені для введення, виведення і тривалого зберігання даних і називаються периферійними.

## **2. Комп'ютерні та телекомунікаційні мережі**

Комп'ютерна мережа – сукупність комп'ютерів і терміналів, з'єднаних за допомогою каналів зв'язку в єдину систему, що задовольняє вимоги розподіленої обробки даних.

У загальному випадку під телекомунікаційною мережею розуміють систему, що складається з об'єктів, які здійснюють функції генерації, перетворення, зберігання і споживання продукту, які називаються вузлами мережі, а також ліній передачі (зв'язку, комунікацій, з'єднань), які здійснюють передачу продукту між пунктами.

Залежно від виду продукту – інформація, енергія, маса - розрізняють відповідно інформаційні, енергетичні і речові мережі.

Інформаційна мережа – комунікаційна мережа, в якій продуктом генерування, переробки, зберігання і використання є інформація. Традиційно для передачі звукової інформації використовуються телефонні мережі, зображень – телебачення, тексту – телеграф (телетайп). У даний час все більшого поширення набувають інформаційні мережі інтегрального обслуговування, що дозволяють передавати в єдиному каналі зв'язку звук, зображення і дані.

Обчислювальна мережа – інформаційна мережа, до складу якої входить обчислювальне обладнання. Компонентами обчислювальної мережі можуть бути ЕОМ і периферійні пристрої, які є джерелами і приймачами даних, переданих по мережі.

Обчислювальні мережі класифікують за рядом ознак:

1. У залежності від відстані між вузлами мережі можна розділити на три класи:

- локальні (LAN, Local Area Network) – охоплюють обмежену територію (зазвичай у межах віддаленості станцій не більше ніж на декілька десятків або сотень метрів одна від одної, рідше на 1 ... 2 км);
- корпоративні (масштабу підприємства) – сукупність пов'язаних між собою локальних мереж, що охоплюють територію, на якій розміщено одне підприємство чи установа в одній або декількох близько розташованих будівлях;
- територіальні – охоплюють значний географічний простір; серед територіальних мереж можна виділити мережі регіональні (MAN, Metropolitan Area Network) і глобальні (WAN, Wide Area Network), що мають відповідно регіональні або глобальні масштаби.

Особливо можна виділити глобальну мережу Інтернет.

2. Важливою ознакою класифікації обчислювальних мереж є їх топологія, яка визначає геометричне розташування основних ресурсів обчислювальних мереж і зв'язків між ними.

Залежно від топології з'єднань вузлів розрізняють мережі шинної (магістральної), кільцевої, зіркової, ієрархічної, довільної структури.

Серед обчислювальних мереж найбільш поширені:

- шинна (bus) – локальна мережа, в якій зв'язок між будь-якими двома станціями встановлюється через один загальний шлях і дані, що передаються будь-якою станцією, одночасно стають доступними для всіх інших станцій, підключених до цього ж середовища передачі даних;
- кільцева (ring) – вузли пов'язані кільцевою лінією передачі (до кожного вузлу підходять лише дві лінії). Дані, проходячи кільцем, по черзі стають доступними всім вузлам мережі;
- зіркова (star) – є центральний вузол, від якого розходяться лінії передачі до кожного з інших вузлів.

Топологічна структура мережі значно впливає на її пропускну здатність, стійкість мережі до відмов її обладнання, на логічні можливості та вартість мережі.

3. Залежно від способу управління розрізняють мережі:

- «клієнт-сервер» – у них виділяється один або декілька вузлів (їх назва – сервери), що виконують у мережі керуючі або спеціальні обслуговуючі функції, а інші вузли (клієнти) є термінальними, у них працюють користувачі. Мережі «клієнт-сервер» розрізняються за характером розподілу функцій між серверами, тобто за типами серверів (наприклад, файл-сервери, сервери баз даних). При спеціалізації серверів за певними додатками маємо мережу розподілених обчислень. Такі мережі відрізняються також від централізованих систем, побудованих на мейн-фреймах;
- однорангові – у яких усі вузли рівні. Оскільки в загальному випадку під клієнтом розуміється об'єкт (пристрій або програма), що запитує деякі послуги, а під сервером – об'єкт, що надає ці послуги, то

кожен вузол в однорангових мережах може виконувати функції і клієнта, і сервера.

4. Залежно від того, однакові або неоднакові ЕОМ застосовують у мережі, розрізняють мережі однотипних ЕОМ, звані однорідними, і різнотипних ЕОМ – неоднорідні або гетерогенні. У великих автоматизованих системах, як правило, мережі виявляються неоднорідними.

5. Залежно від прав власності на мережі, вони можуть бути мережами загального користування (public) або приватними (privat).

Будь-яка комунікаційна мережа повинна включати такі основні компоненти: передавач, повідомлення, засоби передачі, приймач.

Передавач – це пристрій, що є джерелом даних.

Приймач – пристрій, який приймає дані.

Приймачем можуть бути комп'ютер, термінал або будь-який цифровий пристрій.

Повідомлення – цифрові дані певного формату, призначені для передачі. Це може бути файл БД, таблиця, відповідь на запит, текст чи зображення.

Засоби передачі – фізичне середовище та спеціальна апаратура, що забезпечує передачу повідомлень.

Для передачі повідомлень у обчислювальних мережах використовуються різні типи каналів зв'язку. Найбільш поширені виділені телефонні канали і спеціальні канали для передачі цифрової інформації. Застосовуються також радіоканали та канали супутникового зв'язку.

Каналом зв'язку називають фізичне середовище та апаратні засоби, що здійснюють передачу інформації між вузлами комутації.

Потреби формування єдиного світового простору сприяли створенню глобальної мережі Інтернет. У даний час Інтернет залучає користувачів своїми інформаційними ресурсами та сервісами (послугами), якими користується близько мільярда людей у всіх країнах світу. До мережевих послуг відносяться електронні дошки оголошень (Bulletin Board System – BBS), електронна пошта (e-mail), телеконференції або групи новин (News Group), обмін файлами між комп'ютерами (FTR), паралельні бесіди в Інтернеті (Internet Relay Chat - IRC), пошукові системи.

У кожній локальній або корпоративній мережі зазвичай є, принаймні, один комп'ютер, який має постійне підключення до Інтернету за допомогою лінії зв'язку з високою пропускнуною спроможністю (сервер Інтернету). Інтернет надає людині невичерпні можливості пошуку потрібної інформації різного характеру.

Практично всі програми містять, крім довідкової системи, електронну та друковану документацію. Ця документація є джерелом корисної інформації про програму, і нехтувати нею не слід.

Знайомство з програмою починається з інформаційних екранів, які супроводжують її установку. Поки йде установка, слід дізнатися якомога



більше про призначення програми і про її можливості. Це допомагає зрозуміти, що слід розшукувати в програмі після її установки.

Друкована документація додається до програм, куплених в магазинах. Зазвичай це досить великі керівництва об'ємом до декількох сотень сторінок. Саме обсяг такого керівництва часто пригнічує бажання уважно його прочитати. Дійсно, немає сенсу досліджувати керівництво, якщо відповідь на питання можна отримати більш простими засобами. Однак у разі ускладнень, керівництво за програмою - це одне з найбільш зручних джерел необхідної інформації.

У багатьох випадках отримати додаткову допомогу за програмою представляється у вигляді текстових файлів, що входять до складу дистрибутивного комплексу. Історично склалося так, що ці файли зазвичай мають ім'я README, що походить від англійської фрази: «Read me (Прочитай мене)».

Зазвичай файл README містить інформацію про встановлення програми, доповнення та уточнення до друкованого керівництва, а також будь-яку іншу інформацію. Для умовно-безкоштовних програм і невеликих службових програм, які розповсюджуються через Інтернет, цей файл може містити повну електронну версію посібника.

Програми, поширювані через Інтернет, можуть включати й інші текстові інформаційні файли.

У тих випадках, коли ніякі «звичайні» джерела не дозволяють отримати необхідні відомості про програму, можна звернутися до бездонної скарбниці інформації, яку представляє собою Інтернет. Пошук інформації в Інтернеті пов'язаний з деякими труднощами, але зате в мережі є відповіді на будь-які питання.

Усі основні компанії і автори, що виробляють програми для комп'ютерів, представлені в Інтернеті. За допомогою пошукової системи неважко знайти Web-сторінку, присвячену потрібній програмі або серії програм. Така сторінка може містити огляд або короткий опис, відомості про останню версію програми, «заплатки», пов'язані з доопрацюванням програми або виправленням помилок, а також посилання на інші Web-документи, присвячені цим же питанням. Тут же нерідко можна знайти безкоштовні, умовно-безкоштовні, демонстраційні і пробні версії програм.

Мережа Інтернет росте дуже швидкими темпами, і знайти потрібну інформацію серед мільярдів Web-сторінок і файлів стає все складніше. Для пошуку інформації використовуються спеціальні пошукові сервери, які містять більш-менш повну і постійно оновлювану інформацію про Web-сторінки, файли і інші документи, що зберігаються на десятках мільйонів серверів Інтернету.

Різні пошукові сервера можуть використовувати різні механізми пошуку, зберігання та надання користувачеві інформації. Пошукові сервери Інтернету можна розділити на 2 групи:

- спеціалізовані пошукові системи;
- пошукові системи загального призначення [1].

Сучасні пошукові системи часто є інформаційними порталами, які надають користувачам не тільки можливості пошуку документів в Інтернеті, а й доступ до інших інформаційних ресурсів (новин, інформації про погоду, про валютний курс, інтерактивних географічних карт, тощо).

Пошукові системи загального призначення є БД, що містять тематично згруповану інформацію про інформаційні ресурси мережі Інтернет.

Такі пошукові системи дозволяють знаходити Web-сайти або Web-сторінки за ключовими словами в БД або шляхом пошуку в ієрархічній системі каталогів.

Інтерфейс таких пошукових систем загального призначення містить список розділів каталогу та поле пошуку. У поле пошуку користувач може ввести ключові слова для пошуку документа, а в каталозі вибрати певний розділ, що звужує поле пошуку і таким чином прискорює пошук.

Заповнення БД здійснюється за допомогою спеціальних програм-роботів, які періодично «обходять» Web-сервери Інтернету.

Програми-роботи читають всі документи, які зустрічають, виділяють у них ключові слова і заносять в БД, яка містить ЦКБ-адреси документів.

Так як інформація в Інтернеті постійно змінюється (створюються нові Web-сайти і сторінки, видаляються старі, змінюються їх ЦКБ-адреси і так далі), пошукові роботи не завжди встигають відстежити всі ці зміни. Інформація, що зберігається в БД пошукової системи, може відрізнятись від реального стану Інтернету, і тоді користувач в результаті пошуку може отримати адресу вже не існуючого або переміщеного документа.

З метою забезпечення більшої відповідності між змістом БД пошукової системи і реальним станом Інтернету більшість пошукових систем дозволяють автору нового або переміщеного Web-сайту самому внести інформацію в БД, заповнивши реєстраційну анкету. У процесі заповнення анкети розробник сайту вносить ЦКБ-адресу сайту, його назву, короткий опис змісту сайту, а також ключові слова, за якими найлегше буде знайти сайт.

Сайти в БД реєструються за кількістю їх відвідувань у день, тиждень або місяць. Відвідуваність сайтів визначається за допомогою спеціальних лічильників, які можуть бути встановлені на сайті. Лічильники фіксують кожне відвідування сайту і передають інформацію про кількість відвідувань на сервер пошукової системи.

Пошук документа в БД пошукової системи здійснюється за допомогою введення запитів в поле пошуку. Простий запит містить одне або кілька ключових слів, які є головними для цього документа. Можна також використовувати складні запити, що використовують логічні операції, шаблони і так далі.

Спеціалізовані пошукові системи дозволяють шукати інформацію в інших інформаційних «шарах» Інтернету: серверах файлових архівів, поштових серверах, тощо.

### 3. Методи класифікації даних

Класифікація даних на основі логістичної регресії полягає в тому, що простір вихідних значень може бути розділений лінійною кордоном (тобто прямою) на дві області, що відповідають класам.

Логістична регресія схожа на лінійну тим, що в ній теж потрібно знайти значення коефіцієнтів для вхідних змінних. Вихідне значення перетворюється з допомогою нелінійної або логістичної функції.

Значенням функції є ймовірність того, що дане початкове значення належить до певного класу.

Логістична функція виглядає як велика буква S і перетворює будь-яке значення в число в межах від 0 до 1.

Припустимо, що у нас є тільки два класи (для задач з великою кількістю класів – Множинна логістична регресія) і ймовірність, яку ми будемо визначати,  $P+$  ймовірності того, що деяке значення належить класу "+". І звичайно  $P- = 1 - P+$ . Таким чином, результат логістичної регресії завжди знаходиться в інтервалі  $[0, 1]$ .

Кордоном у разі двох вимірів є просто пряма лінія без вигинів. У разі трьох – площина, і так далі. Ця межа задається в залежності від наявних вихідних даних і навчального алгоритму. Щоб все працювало, точки вихідних даних повинні розділятися лінійним кордоном на дві вищезазначених області. Якщо точки вихідних даних задовольняють цій вимозі, то їх можна назвати лінійно розділяємі (рис. 1).

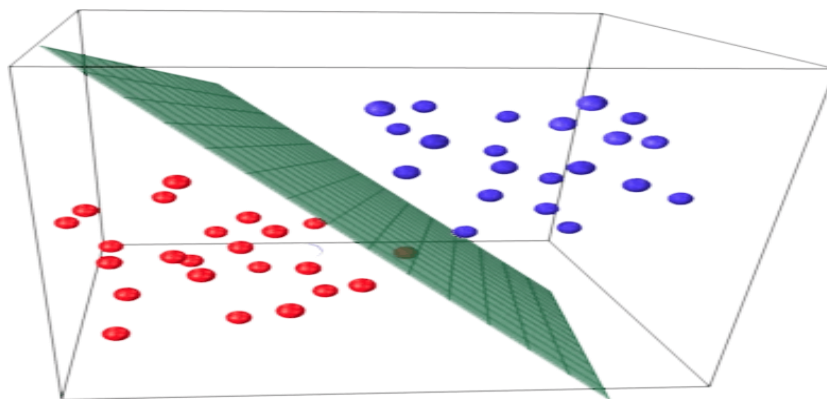


Рис. 1. Дві області, що відповідають класам

Зазначена розділяюча площина називається лінійним дискримінантом, так як вона є лінійною з точки зору своєї функції, і дозволяє моделі проводити розподіл, дискримінацію точок на різні класи.

Якщо неможливо провести лінійний поділ точок у вихідному просторі: перетворити вектори ознак в простір з великою кількістю вимірювань,

додавши додаткові ефекти взаємодії, члени більш високого ступеня та ін. Використання лінійного алгоритму в такому просторі дає певні переваги для навчання нелінійної функції, оскільки межа стає нелінійною при поверненні в початковий простір.

Для наочності представимо нашу базу даних в двомірному вимірі (вік і дохід), у вигляді безлічі об'єктів, що належать класам 1 (помаранчева мітка) і 2 (сіра мітка).

Об'єкти з двох класів (рис. 2):

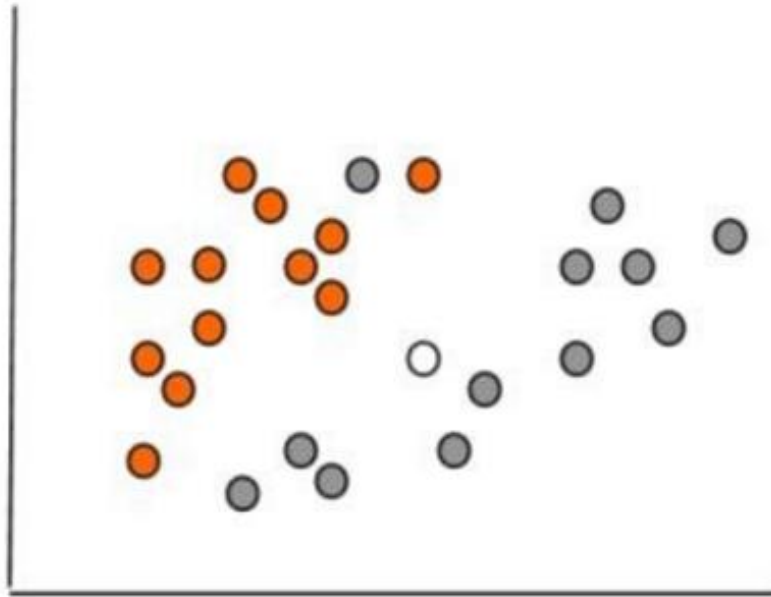


Рис. 2. База даних об'єктів, що належать класам 1 (помаранчева мітка) і 2 (сіра мітка)

Рішення нашої задачі буде полягати в тому, щоб визначити, до якого класу відноситься новий клієнт, на малюнку позначений білою міткою.

Для проведення класифікації за допомогою математичних методів необхідно мати **формальний опис об'єкта**, яким можна оперувати, використовуючи математичний апарат класифікації. Таким описом виступає **база даних**.

Кожен об'єкт (запис бази даних) несе інформацію про деяку властивість об'єкта.

Набір вихідних даних (або вибірку даних) розбивають на два безлічі: навчальне і тестове.

Навчальна множина (training set) – безліч, яке включає дані, використовуються для навчання (конструювання) моделі.

Така безліч містить вхідні та вихідні (цільові) значення прикладів. Вихідні значення призначені для навчання моделі. Тестова (test set) безліч також містить вхідні та вихідні значення прикладів. Тут вихідні значення використовуються для перевірки працездатності моделі.

Процес класифікації складається з двох етапів: конструювання моделі та її використання.

1. Конструювання моделі: опис безлічі визначених класів.

Кожен приклад набору даних відноситься до одного зумовленого класу. На цьому етапі використовується навчальна множина, на ньому відбувається конструювання моделі. Отримана модель представлена класифікаційними правилами, деревом рішень або математичною формулою.

2. Використання моделі: класифікація нових або невідомих значень.

Оцінка правильності (точності) моделі.

1. Відомі значення з тестового прикладу порівнюються з результатами використання отриманої моделі.

2. Рівень точності – відсоток правильно класифікованих прикладів в тестовому безлічі.

3. Тестове безліч, тобто безліч, на якому тестується побудована модель, не повинно залежати від навчальної множини.

Якщо точність моделі допустима, можливе використання моделі для класифікації нових прикладів, клас яких невідомий.

Контрольні запитання.

1. Що таке інформаційний процес?
2. Які є основні способи обробки інформації?
3. Які є властивості у інформаційних систем?
4. Як класифікуються засоби обробки інформації?
5. Які є мережі та чим вони відрізняються?
6. У чому полягає ідея класифікації даних на основі логістичної регресії?
7. Основні етапи класифікації даних.

## Лекція № 4

### Методи статистичного оцінювання і порівняння вибірок. Непараметричні методи перевірки однорідності вибірок

#### План

1. Статистичний аналіз, класифікація методів статистичного аналізу
2. Описова статистика
  - 2.1. Частотний розподіл
  - 2.2. Відсоткові показники
  - 2.3. Заходи центральної тенденції
  - 2.4. Міри розкиду даних
3. Методи вторинної статистичної обробки результатів
  - 3.1. Методи порівняння елементарних статистик (параметричні та непараметричні методи)
  - 3.2. Кореляційний аналіз
4. Роль і значення графічного методу в статистиці

#### **1. Статистичний аналіз, класифікація методів статистичного аналізу**

Статистика – це точна наука, що вивчає методи збору, аналізу і обробки даних, які описують масові дії, явища і процеси. Дані, що вивчаються в статистиці, зачіпають не окремі об'єкти, а їх сукупності.

Статистика є одним з інструментів аналізу експериментальних даних і клінічних спостережень, а також мовою, за допомогою якої повідомляються отримані математичні результати.

Статистичні методи включають як прості методи, які доступні навіть не підготовленим користувачам, так і складні математичні процедури, доступні лише кваліфікованим фахівцям високого класу.

Методами статистичної обробки результатів експерименту називаються математичні прийоми, формули, способи кількісних розрахунків, за допомогою яких показники, що одержані в ході експерименту, можна узагальнювати, приводити в систему, виявляючи приховані в них закономірності.

Головна мета будь-якого статистичного методу – представити кількісні дані в систематизованій і стислій формі з тим, щоб полегшити їх розуміння.

Всі методи статистичного аналізу умовно діляться на первинні і вторинні.

**ПЕРВИННІ:** методи, за допомогою яких можна отримати показники, що безпосередньо відображають результати отриманих в експерименті вимірювань. Це: визначення середньої арифметичної, дисперсії, моди, медіани.

ВТОРИННІ – методи статистичної обробки, за допомогою яких на базі первинних даних виявляють приховані в них статистичні закономірності. Це:

кореляційний аналіз,  
регресійний аналіз,  
факторний аналіз,  
методи порівняння первинних даних двох або декількох вибірок.

Генеральна сукупність і вибірка, члени досліджуваної сукупності наведені на рис. 1.

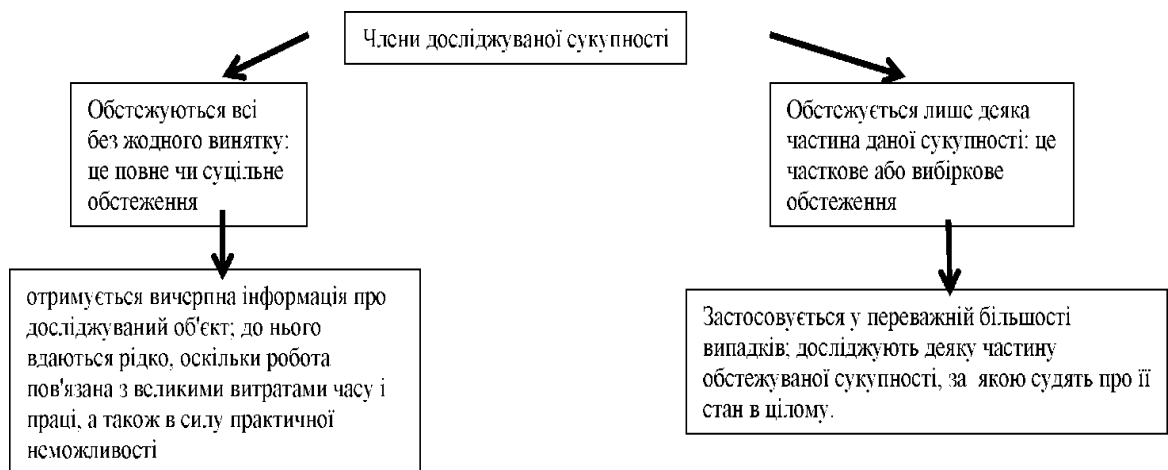


Рис. 1. Члени досліджуваної сукупності

Сукупність, з якої відбирають певну частину її членів для спільного вивчення, називають генеральною.

Відібрана тим чи іншим способом частина генеральної сукупності отримала назву вибіркової сукупності або вибірки.

Загальну суму членів генеральної сукупності називають її обсягом і позначають буквою  $N$ .

Обсяг генеральної сукупності нічим не обмежений, тобто генеральну сукупність представляють як нескінченно велику безліч відносно однорідних одиниць або членів, що складають її зміст. Обсяг вибірки, що позначається буквою  $n$ , може бути і великим, і малим, але він не може містити менше двох одиниць.

## 2. Описова статистика

Описова статистика дозволяє узагальнювати первинні результати, отримані при спостереженні або в експерименті. Процедури тут зводяться:

до угруповання даних по їх значенням,  
побудови розподілу їх частот,  
виявлення центральних тенденцій розподілу (наприклад, середньої арифметичної, моди, медіани),

до оцінки розкиду даних по відношенню до знайденої центральної тенденції.

## 2.1 Частотний розподіл

Як правило, в результаті емпіричного дослідження буває досить багато вихідних первинних даних, які підлягають статистичній обробці. Наприклад, колонка з 1000 тестових показників. Оскільки їх багато – можна скласти таблицю частотного розподілу: Ряд первинних даних виписують у перший стовпчик таблиці в порядку убутання.

Коли показники розподілені по порядку, підраховують кількість випадків для кожного показника. Отримане таким способом число і є частота (кількість випадків) для відповідного показника. Сума всіх частот дорівнює загальному числу тестових показників (або обсягом вибірки  $n$ ).

У другій стовпчик впишемо частоту зустрічаємості кожного первинного результату (рис. 2).

| Первинний результат | Частота |
|---------------------|---------|
| 26                  | 2       |
| 25                  | 2       |
| 24                  | 8       |
| 23                  | 5       |
| 22                  | 6       |
| 21                  | 7       |
| 20                  | 9       |
| 19                  | 5       |
| 18                  | 4       |
| 17                  | 1       |
| 16                  | 2       |
| 15                  | 3       |
| 14                  | 2       |

Рис. 2. Таблиця частотного розподілу даних по тесту

Інформація, що міститься в частотному розподілі, може бути також представлена графічно у вигляді кривої розподілу (рис. 3).





Рис. 3. Крива розподілу

Для того, щоб зробити узагальнені дані про характер розподілення результатів по тесту і в разі, якщо отримано занадто велике число значень первинного результату, необхідно зробити угруповання даних і провести аналогічну процедуру побудови частотного розподілу. Згрупуємо дані, представлені в табл. (рис. 2), об'єднуючи їх по 3 одиниці в кожній групі (або з інтервалом 3 одиниці). Угруповання даних проводиться від мінімального значення до максимального. Для того, щоб інтервали значень були рівномірними, додамо ще два значення в нижній частині ряду (12 і 13), так як 26 балів є максимальним для даної методики. Частота цих первинних показників, що зустрічаються в нашій вибірці дорівнює 0. Частотний розподіл придбає такий вигляд (рис. 4):

| Інтервал значень | Частота |
|------------------|---------|
| 24-26            | 12      |
| 21-23            | 18      |
| 18-20            | 18      |
| 15-17            | 6       |
| 12-14            | 2       |

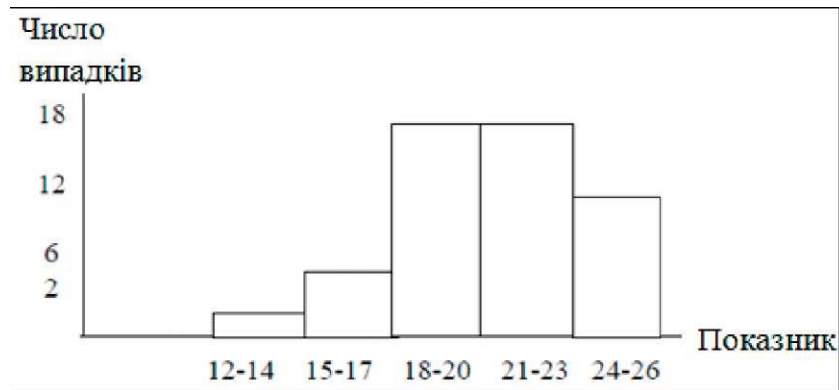


Рис. 4. Частотний розподіл

Судячи з характеру розподілу, представленого на рис. 4, воно не є нормальним і характеризується невеликою асиметрією зі зрушенням в бік високих значень.

Ідеальна нормальна крива зображена на рисунку 5:

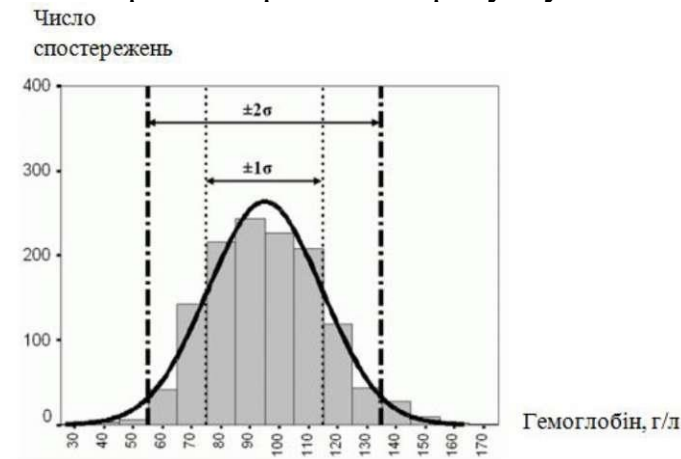


Рис. 5. Ідеальна нормальна крива

Цей тип кривої володіє важливими математичними властивостями і на ній засновані багато видів статистичного аналізу. По суті, ця крива означає, що число випадків максимально в середині розподілу і поступово спадає до її країв. Крива симетрична і має єдиний пік в центрі.

Більшість розподілів чисельних показників наближаються до нормальної кривої. Можна відзначити таку закономірність: чим більше група, тим ближче розподіл показників до нормальної кривої

## 2.2. Відсоткові показники

Відсоткові показники використовуються для того, щоб частотний розподіл за тією чи іншою змінною привести до основи 100 (аналогічно, пропорції використовуються для приведення даних до основи 1). У такому

вигляді дані є кращими в інтуїтивному сенсі в порівнянні з «сирим» частотним розподілом.

Приклад (успішність здачі сесії):

| Успішність                        | Частота | %     | Пропорції |
|-----------------------------------|---------|-------|-----------|
| Були трійки                       | 28      | 45,2  | 0,452     |
| Без трійок, в основному на чотири | 11      | 17,7  | 0,177     |
| Без трійок, в основному на п'ять  | 13      | 21,0  | 0,210     |
| На відмінно                       | 10      | 16,1  | 0,161     |
| Разом                             | 62      | 100,0 | 1,000     |

### 2.3. Заходи центральної тенденції

Заходи центральної тенденції (мода, медіана і середнє арифметичне) дають інформацію про типові або центральні значення розподілу.

Мода говорить про значення, що найбільш часто зустрічається, Медіана – про серединне значення,

Середнє арифметичне – про найбільш очікуване значення.

Найбільш часто використовують середнє арифметичне. Його обчислюють, розділивши суму всіх значень даних на число цих даних.

### 2.4 Міри розкиду даних

Для більш повного опису результатів емпіричного дослідження використовуються міри розкиду (або варіативності) даних, що характеризують ступінь індивідуальних відхилень від центральної тенденції. Найбільш наочним і відомим способом подання розкиду є розмах розподілу, тобто різницю між найвищим і найнижчим результатом. Але ця міра вкрай неточна і нестійка, тому що вона визначається тільки двома показниками, і єдиний надзвичайно високий або низький результат може помітно вплинути на величину розмаху. Більш точний метод вимірювання розкиду даних заснований на обліку різниці між кожним індивідуальним результатом і середньоарифметичним значенням по групі. Такий мірою розкиду є дисперсія або середній квадрат відхилення ( $\sigma$ ).

Дисперсія як статистична величина характеризує, наскільки приватні значення відхиляються від середньої величини в даній вибірці. Чим більша дисперсія, тим більше відхилення або розкид даних.

Дуже часто замість дисперсії для виявлення розкиду приватних даних щодо середньої використовують похідну від дисперсії величину – стандартне (або вибіркове) відхилення. Воно дорівнює квадратному кореню, який витягується з дисперсії, і позначається тим же знаком, тільки без квадрата ( $\sigma$ ). Ця величина в ряді випадків виявляється більш зручною характеристикою варіювання, ніж дисперсія, так як виражається в тих же одиницях, що і середня арифметична величина. Визначають стандартне відхилення від середнього за формулою:

$$S_{\bar{x}} = \sqrt{\frac{\Delta x_1^2 + \Delta x_2^2 + \dots + \Delta x_n^2}{n}}$$

Для того, щоб більш точно оцінити стандартне відхилення для малих вибірок (з числом елементів менше 30), знаменнику виразу під коренем треба використовувати не  $n$ , а  $n-1$ :

$$S_{\bar{x}} = \sqrt{\frac{\Delta x_1^2 + \Delta x_2^2 + \dots + \Delta x_n^2}{n-1}}$$

Особливо чіткою виявляється інтерпретація  $s$  стосовно нормальної або до приблизно нормальної кривої розподілу, тому що тут існує пряма відповідність між  $s$  і відносною кількістю випадків. На рис. 6 по горизонтальній осі відкладені інтервали, які відповідні відхиленню в  $1\sigma$ ,  $2\sigma$  і  $3\sigma$  вправо і вліво від середнього значення ( $M$ ). Відсоток випадків, що припадають на інтервал  $M+1\sigma$  в нормальному розподілі дорівнює 34,13. Оскільки крива симетрична, 34,13% випадків припадає також на інтервали від  $M$  до  $-1\sigma$ , так що діапазон від  $-1\sigma$  до  $+1\sigma$  охоплює 68,26% випадків. Майже всі випадки (99,72%), тобто майже всі показники лежать від  $-3\sigma$  до  $+3\sigma$  щодо середнього значення.

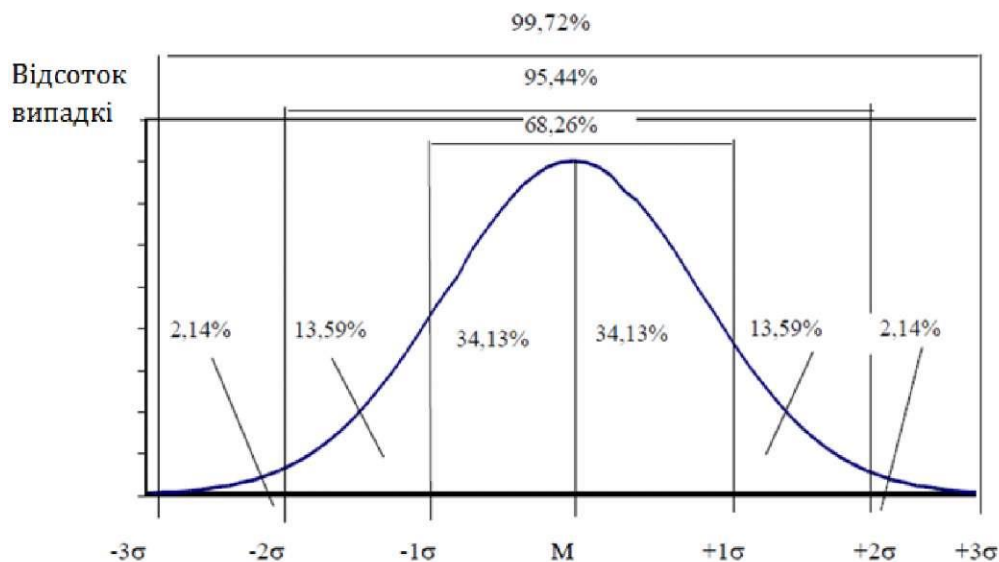


Рис. 6. Приблизно нормальна крива розподілу

Більшість розподілів первинних результатів дослідження ближче до нормального розподілу, ніж до будь-якого іншого.

Властивості нормального розподілу

1. У нормальному розподілі всі міри центральної тенденції рівні між собою, тобто сходяться в одній точці на графіку ( $M = Me = Mo$ ).
2. У нормальному розподілі приблизно 99% всіх значень досліджуваної змінної знаходиться в межі  $M \pm 3\sigma$ . Відсотковий розподіл

випадків на кривій нормального розподілу називають "законом трьох сігм" (рис. 6, 7).

3. Крива нормального розподілу має вигляд дзвона, вона симетрична (асиметрія відсутня,  $A_e = 0$ ) і не має надто гострою або занадто плоскої вершини (ексцес відсутній,  $E_x = 0$ ).

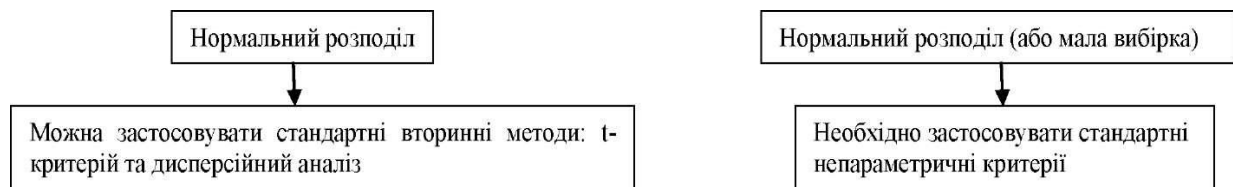


Рис. 7. Застосування стандартних методів та критеріїв

### 3. Методи вторинної статистичної обробки результатів

За допомогою вторинних методів статистичної обробки даних безпосередньо перевіряються, доводяться або спростовуються гіпотези, пов'язані з емпіричним дослідженням.

Найчастіше в дослідженнях застосовують такі методи вторинної статистичної обробки результатів:

- 1) Методи порівняння двох або кількох елементарних статистик (середніх, дисперсій, тощо), що відносяться до різних вибірок;
- 2) Методи встановлення статистичних зв'язків між змінними (наприклад, їх кореляції між собою);
- 3) Методи виявлення внутрішньої статистичної структури емпіричних даних (наприклад, факторний аналіз).

1) Методи порівняння елементарних статистик Жодне дослідження не обходиться без порівнянь. Порівнюють:

- результати, отримані двома різними групами досліджуваних;
- результати, отримані однією вибіркою випробовуваних, але в різний час або в різних умовах.

Про відмінності між ними судять зазвичай за різницею між середніми, дисперсіями і іншими вибірковими показниками. У статистиці широке застосування отримала так звана нульова гіпотеза ( $H_0$ ).

Суть її зводиться до припущення, що різниця між генеральними параметрами порівнюваних груп дорівнює нулю і що відмінності, які спостерігаються між вибірковими характеристиками носять не систематичний, а виключно випадковий характер.

Для перевірки цієї гіпотези використовують спеціальні критерії достовірності (параметричні критерії): 1 – критерій Стюдента і Б – критерій Фішера (за умови нормального розподілу досліджуваної змінної). Перший використовують для порівняльної оцінки середніх величин, другий – для порівняльної оцінки дисперсій.

Якщо розподіл не нормальний, то використовують непараметричні критерії.

## ПАРАМЕТРИЧНІ СТАТИСТИЧНІ МЕТОДИ

### t – критерій Стьюдента

Використовують для порівняння вибірових середніх величин, що належить до двох сукупностей даних, і для вирішення питання про те, чи відрізняються середні значення статистично достовірно одне від іншого.

Метод Стьюдента різний для незалежних і залежних вибірок. Незалежні вибірки отримують при дослідженні двох різних груп випробовуваних. Для аналізу різниці середніх застосовують формулу:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{m_1^2 + m_2^2}}$$

Після того як за допомогою формули вираховано показник t, за спеціальною таблицею для заданого числа ступенів свободи, що дорівнює  $(n_1 + n_2) - 2$  і обраної ймовірності припустимої помилки (таблицю), знаходять потрібне табличне значення t і порівнюють з ним розраховане значення t. Якщо обчислене значення t більше або дорівнює табличному, то роблять висновок про те, що порівнювані середні значення з двох вибірок дійсно статистично достовірно різняться з прийнятою ймовірністю допустимої помилки, тобто нульова гіпотеза не вірна.

До залежних вибірок відносяться, наприклад, результати однієї і тієї ж групи випробовуваних до і після впливу незалежної змінної.

Даний метод порівняння середніх величин застосовується тоді, коли необхідно, наприклад, встановити, вдався чи не вдався експеримент, надав або не чинив він вплив на рівень, наприклад, збільшення сполук чи клітин в організмі. Оцінюються залежні змінні на початку і в кінці експериментального дослідження.

Отримавши такі оцінки і обчисливши середнє по всій вивченій вибірці випробовуваних, ми можемо скористатися критерієм Стьюдента для точного встановлення наявності або відсутності статистично достовірних відмінностей між середніми до і після експерименту.

Якщо виявиться, що вони дійсно вірогідно розрізняються, то можна буде зробити певний висновок про те, що експеримент вдався. В іншому випадку немає переконливих підстав для такого висновку навіть в тому випадку, якщо самі середні величини на початку і в кінці експерименту за своїми абсолютними значеннями різні.

Для визначення достовірності різниці середніх у разі залежних вибірок застосовується наступна формула:

$$t = \frac{\sum d}{\sqrt{\frac{n \sum d^2 - (\sum d)^2}{n-1}}}$$

### **Дисперсійний аналіз (ANOVA)**

На відміну від 1-критерію дисперсійний аналіз (ANOVA) дозволяє порівнювати середні значення трьох і більше груп.

В основі дисперсійного аналізу лежить припущення про те, що одні змінні можуть розглядатися як причини (фактори, незалежні змінні), а інші як слідства (залежні змінні).

Основна мета: дослідження значущості відмінності між середніми за допомогою порівняння дисперсій. Поділ загальної дисперсії на кілька джерел, дозволяє порівняти дисперсію, викликану відмінностями між групами, з дисперсією, викликану внутрішньогруповою мінливістю.

Вибірки (3 і більше) можуть бути як рівними, так і нерівними за чисельністю, як пов'язаними, так і непов'язаними. За кількістю виявлених регульованих факторів дисперсійний аналіз може бути однофакторний (при цьому вивчається вплив одного фактора на результати експерименту), двофакторний (при вивченні впливу двох чинників) і багатофакторним (дозволяє оцінити не тільки вплив кожного з факторів окремо, але і їх взаємодію).

### **НЕПАРАМЕТРИЧНІ СТАТИСТИЧНІ МЕТОДИ**

Непараметричні методи дозволяють обробляти дані "низької якості" з вибірок малого обсягу зі змінними, про розподіл яких мало що або взагалі нічого невідомо.

Непараметричні методи не ґрунтуються на оцінці параметрів (таких як середнє або стандартне відхилення) при описі вибіркового розподілу величини, що цікавить. Тому ці методи іноді також називаються вільними від параметрів або вільно розподіленими.

По суті, для кожного параметричного критерію є, принаймні, один непараметричний аналог. Ці критерії можна віднести до однієї з наступних груп:

- А) критерії відмінності між незалежними вибірками;
- Б) критерії відмінності між залежними вибірками;
- В) критерії залежності між змінними.

#### **А) Критерії відмінності між незалежними вибірками**

Непараметричними альтернативами параметричного критерію для двох незалежних груп є:

- 1) U критерій Манна-Уїтні;
- 2) Критерій серій Вальда-Вольфовиця;
- 3) Двовибірковий критерій Колмогорова-Смірнова.

Кілька незалежних груп: якщо ви маєте кілька груп, то можете використовувати Дисперсійний аналіз (ANOVA). Його непараметричними аналогами є:

- 1) Рангові дисперсійний аналіз Краскела-Уолліса;
- 2) Медіанний тест.

**Б) Критерії відмінності між залежними вибірками Дві залежні вибірки: критерій Вілкоксона і ін.**

Якщо треба порівняти дві змінні, що відносяться до однієї і тієї ж вибірки (наприклад, математичні успіхи студентів на початку і в кінці семестру), то зазвичай використовується критерій для залежних вибірок. Альтернативними непараметричними тестами є:

- 1) Критерій Вілкоксона парних порівнянь;
- 2) Критерій знаків Кілька залежних вибірок.

Якщо розглядається більше двох змінних, що відносяться до однієї і тієї ж вибірки, то зазвичай використовується Дисперсійний аналіз (ANOVA) з повторними вимірами. Альтернативним непараметричним методом є:

- 1) ранговий дисперсійний аналіз Фрідмана;
- 2) критерій Кохрена.

## **2) Кореляційний аналіз**

При вивченні кореляцій намагаються встановити, чи існує якийсь зв'язок між двома показниками в одній вибірці (наприклад, між зростанням і вагою дітей або між рівнем ір і шкільною успішністю) або між двома різними вибірками (наприклад, при порівнянні пар близнюків), і якщо цей зв'язок існує, то чи супроводжується збільшення одного показника зростанням (позитивна кореляція) або зменшенням (негативна кореляція) іншого.

Можна використовувати два різні способи кореляційного аналізу: параметричний метод розрахунку коефіцієнта Брауна - Пірсона ( $r$ ) і обчислення коефіцієнта кореляції рангів Спірмена ( $r_s$ ), який є непараметричним.

Коефіцієнт кореляції (позначається маленькою літерою  $r$ ) і показує нам дві речі: 1) ступінь зв'язку двох змінних і 2) напрямок цього зв'язку (прямий або зворотний зв'язок).

Коефіцієнт кореляції – це величина, яка може варіювати в межах від +1 до -1. У разі повної позитивної кореляції цей коефіцієнт дорівнює плюс 1, а при повній негативній – мінус 1. На графіку (рис. 8) цьому відповідає пряма лінія, що проходить через точки перетину значень кожної пари даних:

---



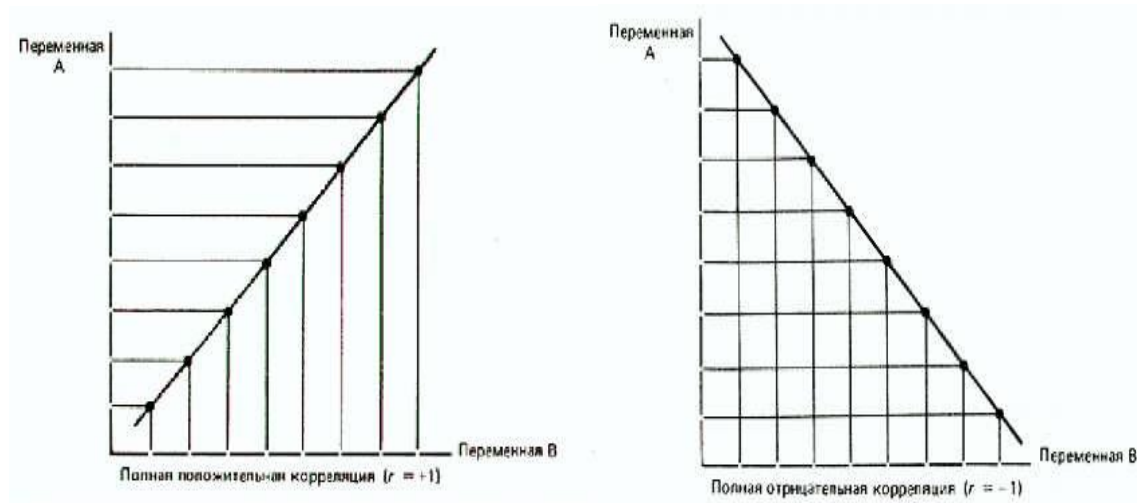


Рис. 8. Позитивний та негативний коефіцієнт кореляції

У разі ж якщо ці точки не шикуються по прямій лінії, а утворюють «хмару», коефіцієнт кореляції за абсолютною величиною стає менше за одиницю і за мірою округлення цієї хмари наближається до нуля (рис. 9):

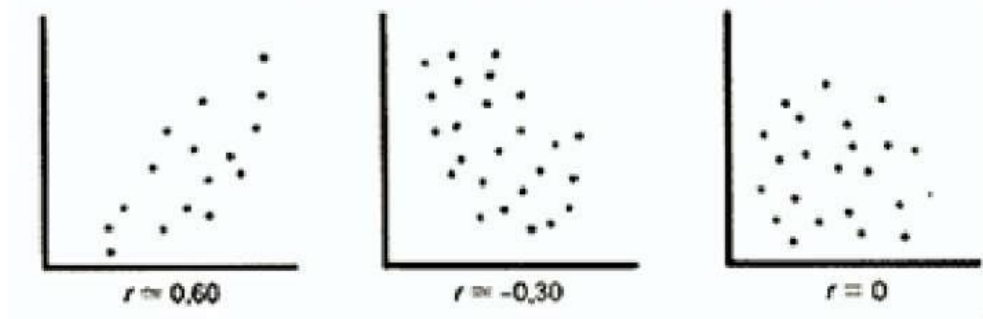


Рис. 9. Коефіцієнт кореляції за абсолютною величиною менше за одиницю

У разі якщо коефіцієнт кореляції дорівнює 0, обидві змінні повністю незалежні одна від одної.

При оцінці сили зв'язку коефіцієнтів кореляції використовується шкала Чеддока. При негативній кореляції значення сили зв'язку між змінними змінюють на протилежні. Таблиця аналізу сили зв'язку між змінними:

| Значення      | Інтерпретація |
|---------------|---------------|
| от 0 до 0,3   | дуже слабкий  |
| от 0,3 до 0,5 | слабкий       |
| от 0,5 до 0,7 | середній      |
| от 0,7 до 0,9 | високий       |
| от 0,9 до 1   | дуже високий  |

Наприклад:

- якщо величина коефіцієнта кореляції між змінними дорівнює  $-0,36$ , то це слабка негативна кореляція, і швидше за все ми не будемо приймати її до уваги;

- якщо величина коефіцієнта кореляції дорівнює  $0$  - змінні не пов'язані між собою;

- якщо величина коефіцієнта кореляції між змінними дорівнює  $0,25$  то це дуже слабка кореляція і в більшості випадків ми не беремо її до уваги;

- якщо величина коефіцієнта кореляції між змінними дорівнює  $0,75$  то це висока кореляція і в своїх інтерпретаціях нам варто звернути на неї увагу;

- якщо величина коефіцієнта кореляції дорівнює  $1$  - змінні повністю взаємопов'язані.

Однак для того, щоб можна було робити висновки про зв'язки між змінними, велике значення має обсяг вибірки: чим вибірка більше, тим вірогідніше величина отриманого коефіцієнта кореляції. Існують таблиці з критичними значеннями коефіцієнта кореляції Браує-Пірсона та Спірмена для різного числа ступенів свободи.

Коефіцієнт кореляції Браує-Пірсона ( $r$ ) – це параметричний показник, для обчислення якого порівнюють середні і стандартні відхилення результатів двох вимірювань.

Коефіцієнт кореляції рангів Спірмена ( $r$ ) – це непараметричний показник, за допомогою якого намагаються виявити зв'язок між рангами відповідних величин в двох рядах вимірів. Цей коефіцієнт розраховувати простіше, проте результати виходять менш точними.

#### **4. Роль і значення графічного методу в статистиці**

Графіком в статистиці називається умовне зображення статистичних даних у вигляді різних геометричних образів: точок, ліній, фігур, тощо. Головна перевага графіків – наочність.

Графіки в статистиці, як правило, використовуються для широкої популяризації даних і полегшення їх сприйняття неспеціалістами, тому в доповідях, промовах і повідомленнях використання статистичних даних часто здійснюється за допомогою графіків. Графіки широко використовуються для узагальнення і аналізу статистичних даних.

Загальні правила побудови графічного зображення

При побудові графіка важливо знайти такі способи зображення, які найкращим чином відповідають змісту і логічній природі зображуваних показників.

Кожен графік складається з графічного образу і допоміжних елементів.

Графічний образ (основа графіка) – це геометричні знаки, тобто сукупність точок, ліній, фігур, за допомогою яких зображуються статистичні показники. Важливо правильно вибрати графічний образ, який повинен відповідати меті графіка і сприяти найбільшій виразності

зображуваних статистичних даних. Графічний образ може являти собою ряд стовпчиків або квадратів і т.п.

Допоміжні елементи уможливають читання графіка, його розуміння і використання. До них відносяться: заголовок графіка, підписи, пояснення.

За характером графічного образу розрізняють графіки об'ємні, лінійні і площинні.

Контрольні запитання.

1. Що таке статистичний аналіз?
2. Яка існує класифікація методів статистичного аналізу?
3. В чому сутність описової статистики?
4. Що таке частотний розподіл?
5. Які існують відсоткові показники?
6. Які є заходи центральної тенденції?
7. Що таке міри розкиду даних?
8. Які є методи вторинної статистичної обробки результатів?
9. Які є методи порівняння елементарних статистик (параметричні та непараметричні методи)?
10. В чому сутність кореляційного аналізу?

## Лекція № 5

### Дисперсійний аналіз. Методи обробки рангових даних.

У багатьох практичних завданнях вплив деяких величин на вихідну величину об'єкта неможливо оцінити кількісно. Однак дослідника цікавить питання, наскільки суттєвий вплив того чи іншого фактора (або їх комбінації) на величини, що розглядаються.

Нехай, наприклад, яка-небудь технологічна операція виконується паралельно на кількох роботизованих лініях. Для правильної організації подальших етапів технологічного процесу треба знати, в якій мірі однотипними є середні розміри деталей, що одержані на паралельно працюючих роботизованих лініях. При експериментальних дослідженнях, які проводяться дослідниками на різних лініях, важливо вивчити вплив двох факторів на результат експерименту – дослідника і устаткування. Якщо ж до того дослідження проводились в різний час (або в різних місцях), то вводиться один фактор – час (місце) проведення експерименту. Аналогічне завдання виникає при дослідженні партій виробів, які одержані від різних постачальників при з'ясуванні впливу різних якостей сировини на якість продукції і т. ін.

Нижче будуть розглянуті задачі, пов'язані з експериментальними дослідженнями, зокрема з перевіркою правильності їх організації. У загальному випадку задача виглядає таким чином.

Нехай:

- 1) вихідна величина (ознака, відгук) внаслідок фізичних властивостей залежить від  $n$  факторів, які не мають кількісного опису, від їх парних взаємодій;
- 2) кожний фактор можна варіювати на декількох рівнях (експеримент проводять кілька операторів, застосовуються різні методи вимірювання і т. ін.);
- 3) кожен дослідну операцію можна спостерігати декілька разів, тобто реалізується серія паралельних дослідів.

Треба визначити, в якій мірі на вихідну величину (на фон впливу випадкових величин) впливають дані якісні фактори, зробити їх співвідношення і ранжувати. При цьому вважають, що відгук (ознака) в загальному випадку є випадковою величиною, розподіленою за нормальним законом, дисперсія у всіх дослідів однорідна, тобто накладається умова відтворюваності дослідів. Ці припущення у процесі проведення дослідження необхідно перевірити.

#### 1. Однофакторний дисперсійний аналіз

Розглянемо ідеальний варіант – випадкові впливи відсутні, необхідно дослідити вплив тільки одного фактора. Найкращою оцінкою впливу

фактора може служити величина, аналогічна дисперсії, що характеризує розсіювання вихідної величини  $y_i \sim$  біля деякого середнього значення. Проте не треба забувати, що в даному разі спостерігається тільки аналогія з дисперсіями, оскільки  $\tilde{y}_i$  - детерміновані величини, такі величини, які впливають і похибки сприйняття відсутні.

Тоді кількісна оцінка впливу фактора може бути подана у вигляді

$$\hat{\sigma}_x^2 = \sum_{i=1}^N (\tilde{y}_i - \bar{y})^2 / N - 1,$$

де  $N$  – кількість приладів (рівнів), кожним з яких проводиться одне вимірювання.

У дійсності спостерігається вплив випадкових неврахованих величин, які в сукупності можна позначити  $\varepsilon$ . Вплив їх представимо у вигляді  $\varepsilon^2$ . При цьому дисперсія вихідної величини вже буде визначатися не тільки впливом фактора  $X$ , але й випадковою величиною  $\varepsilon$ . Співставлення впливу цих величин через співставлення зумовлених ними дисперсій і є основою дисперсійного аналізу. Якщо досліджуваний фактор несуттєвий і дисперсія відтворюваності, яка характеризує один дослід, відома, то загальна дисперсія буде в основному визначатися дисперсією відтворюваності.

Якщо ж фактор суттєвий, то можна вважати, що

$$S_0^2\{y\} = \sigma_x^2 + \sigma_\varepsilon^2.$$

Виходячи з даного співвідношення, можна знайти дисперсію, яку зумовлює фактор, що вивчається. Звичайно дисперсія відтворюваності  $\varepsilon^2$  невідома, і її треба знайти за результатами експериментальних досліджень, тому проводять паралельні досліді для кожного рівня варіювання факторів.

Вважатимемо, що дослідження об'єкта проводиться одночасно різними приладами з метою зменшення впливу фактора – приладу на результат дослідження (похибка показань). З'ясуємо, чи можна систематичні похибки приладів вважати однаковими.

Нехай число приладів буде  $N$  (фактор варіюється на  $N$  рівнях) і для кожного приладу (рівня) проводиться серія з  $m$  паралельних дослідів. Число дослідів при реалізації однофакторного дисперсійного аналізу буде  $N \cdot m$ .

Результати дослідів наведені в табл. 1.

Таблиця 1

| i | l=1              | l=2              | ... | l=m              | $\bar{y}_i$   |
|---|------------------|------------------|-----|------------------|---------------|
| 1 | $\tilde{y}_{11}$ | $\tilde{y}_{12}$ | ... | $\tilde{y}_{1m}$ | $\tilde{y}_1$ |
| 2 | $\tilde{y}_{21}$ | $\tilde{y}_{22}$ | ... | $\tilde{y}_{2m}$ | $\tilde{y}_2$ |
| · |                  |                  |     |                  |               |
| · | ...              | ...              | ... | ...              | ...           |
| · |                  |                  |     |                  |               |
| N | $\tilde{y}_{N1}$ | $\tilde{y}_{N2}$ | ... | $\tilde{y}_{Nm}$ | $\tilde{y}_N$ |

За результатами дослідження для кожного i-го рівня незалежної змінної (різновиднос ті приладу) знаходять середнє значення (вважаємо однакову кратність проведення дослідів):

$$\bar{y}_i = \sum_{l=1}^m \frac{y_{il}}{m}$$

Розкид значень відгуків у фіксованому рядку (для конкретного приладу) визначається сукупною дією випадкових величин і характеризується оцінкою дисперсії відтворюваності. Розкид між середніми значеннями вихідних величин визначається впливом фактора. Якщо систематичні похибки приладів однакові, то треба чекати підвищеного розсіювання, вибіркового засобів середніх  $\bar{y}_i$ .

Незміщена оцінка дисперсії відтворюваності для всієї сукупності дослідів визначиться у вигляді

$$S_0^2\{y\} = \sum_{i=1}^N \sum_{l=1}^m (\tilde{y}_{il} - \bar{y})^2 / (Nm - 1)$$

$$\bar{y} = \sum_{i=1}^N y_i / N$$

Розглянемо чисельник даного виразу  $Q_0^2$ , вводячи під дужки  $-\bar{y}_i i + \bar{y}_i$

$$Q_0^2 = \sum_{i=1}^N \sum_{l=1}^m [(\tilde{y}_{il} - \bar{y})^2 + (\bar{y}_i - \bar{y})^2] = \sum_{i=1}^N \sum_{l=1}^m [\tilde{y}_{il} - \bar{y}_i]^2 + 2 \sum_{i=1}^N \sum_{l=1}^m (\tilde{y}_{il} - \bar{y}_i) + (\bar{y}_i - \bar{y}) + \sum_{i=1}^N \sum_{l=1}^m (\bar{y}_i - \bar{y})^2$$

Розглянемо окремо другу складову в правій частині виразу  $Q_0^2$

$$\sum_{i=1}^N \sum_{l=1}^m (\tilde{y}_{il} - \bar{y}_i) + (\bar{y}_i - \bar{y}) = \sum_{i=1}^N (\bar{y}_i - \bar{y}) \sum_{l=1}^m (\tilde{y}_{il} - \bar{y}_i)$$

Оскільки сума відхилень від середнього в i – й серії дорівнює нулю

$$\sum_{l=1}^m (\tilde{y}_{il} - \bar{y}_i) = 0$$

то і друга складова також дорівнює нулю.

Таким чином, повна сума квадратів відхилень окремих спостережень від  $\bar{y}$ :

$$Q_0^2 = \sum_{i=1}^N \sum_{l=1}^m [(\tilde{y}_{il} - \bar{y}_i)^2 + \sum_{i=1}^N \sum_{l=1}^m (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^N \sum_{l=1}^m (\tilde{y}_{il} - \bar{y}_i)^2 + m \sum_{i=1}^N (\bar{y}_i - \bar{y})^2$$

або в скороченому вигляді:  $Q_0^2 = Q_\varepsilon^2 = Q_x^2$

$$Q_0^2 = \sum_{i=1}^N \sum_{l=1}^m [(\tilde{y}_{il} - \bar{y})^2;$$

$$Q_\varepsilon^2 = \sum_{i=1}^N \sum_{l=1}^m (\tilde{y}_{il} - \bar{y}_i)^2;$$

$$Q_x^2 = m \sum_{i=1}^N (\bar{y}_i - \bar{y})^2$$

Складова  $Q_x^2$  являє собою суму квадратів різниць між середніми окремих серій (рядків, прикладів) і загального середнього у за всією сукупністю спостережень і характеризує ступень розходження систематичних похибок в окремих прикладах. Її ще називають «розсіянням за факторами».

Складова  $Q_\varepsilon^2$  являє собою суму квадратів різниць між окремими спостереженнями і середнім відповідним серії  $\bar{y}_i$  (середнє значення показів даного прикладу) і характеризує «залишкове розсіювання» випадкових похибок дослідів.

Таким чином, повне розсіювання показів приладів  $Q_0^2$  складається із двох компонент, характеризуючи розсіювання між приладами, тобто різниці між їх систематичними похибками  $Q_x^2$  і розсіяння «в середині» приладів (серій), яке характеризує однакову (на основі передумов дисперсійного аналізу) для всіх приладів варіацію під дією випадкових величин  $Q_\varepsilon^2$ .

Вважаємо, що гіпотеза рівнос ті систематичних похибок правильна, тому нормальні розподіли для всіх приладів тотожні, тобто похибки мають однаковий центр розподілу (систематична похибка) і дисперсію  $\sigma_x^2$ . У цьому випадку всі  $Nm$  спостережень можна розглядати як вибірку із однієї і тієї ж нормальної сукупності, а  $Q_0^2 / (Nm - 1)$ , як вже відмічалось, є незміщеною оцінкою дисперсії  $2\sigma_x^2$  х за цією вибіркою.

Звідси виходить, що відношення  $Q_0^2 / \sigma_\varepsilon^2$  буде відповідати розподілу  $\chi^2$  з  $(Nm-1)$  ступенями свободи.

Проте середні за групами (приладами)  $\bar{y}_i$  також згідно з припущенням нормально розподілені з дисперсією  $|\sigma_x^2 / m|$  кожна, і незалежні одна від одної.

Тому

$$\sum_{i=1}^N (\bar{y}_i - \bar{y})^2 / (N - 1) = Q_x^2 / m(N - 1)$$

є незміщеною вибірковою характеристикою дисперсії  $|\sigma_x^2 / m|$ , одержаної на основі N спостережень величини  $\bar{y}_i$ . У результаті можна зробити висновок, що величина

$$\frac{\frac{Q_x^2}{m}}{\frac{\sigma_\varepsilon^2}{m}} = \frac{Q_x^2}{Q_\varepsilon^2} = \sum_{i=1}^N \sum_{l=1}^m (\bar{y}_i - \bar{y})^2 / \sigma_\varepsilon^2$$

розподілена за законом  $X^2$  з (N-1) ступенями свободи.

Сума квадратів відхилень від середнього в кожній серії, віднесена до дисперсії

$$\sum_{l=1}^m (\bar{y}_{il} - \bar{y}_i)^2 / \sigma_\varepsilon^2$$

також розподілена за законом  $X^2$  з (m-1) ступенями свободи.

Відповідні властивості композиції для N серій (приладів), компонента

$$Q_\varepsilon^2 / \sigma_\varepsilon^2 = \sum_{i=1}^N \sum_{l=1}^m (\bar{y}_{il} - \bar{y}_i)^2 / \sigma_\varepsilon^2$$

також розподілена за законом  $X^2$  з N(m-1) ступенями свободи. Таким чином,

$Q_\varepsilon^2 / N(m - 1)$  є також оцінкою параметра  $Q_\varepsilon^2$

Із викладеного випливає, що у разі рівності систематичних похибок приладів (несуттєвості впливу фактора, що досліджується) існують три незміщені оцінки  $Q_\varepsilon^2$ . Відношення двох однорідних оцінок дисперсій

$$F_p = [Q_x^2 / (N - 1)] / [Q_\varepsilon^2 / N(m - 1)] \quad (1)$$

відповідатиме F - розподілу з (N-1) і N(m-1) ступенями свободи. Задаючись  $\alpha$  рівнем значущості, на основі таблиці F- розподілу можна встановити відповідну  $\alpha$  границю, так що

$$P(F_p > F_\alpha) = \alpha / 100$$

Розглянемо випадок, коли фактор, що досліджується, суттєвий, тобто гіпотеза про рівність систематичних похибок (центрів розподілу випадкових похибок) невірна, але параметр  $Q_\varepsilon^2$  у всіх N сукупностях один і той же. Зміна центрів рядків розподілів, тобто зміна  $\tilde{y}_{il}$  на  $\tilde{y}_{il} - c_i$  (де  $c_i$  – систематична складова похибки i-го приладу) не змінить значення  $Q_\varepsilon^2$  що, як і раніше, розподілено за законом  $X^2$  з N(m-1) ступенями свободи, а  $Q_\varepsilon^2 / N(m - 1)$  залишається незміщеною оцінкою  $Q_\varepsilon^2$ .

Проте чисельник виразу (1) саме враховує розходження між центрами розподілу  $c_i$  і має тенденцію до збільшення при збільшенні розходження



між систематичними складовими. Тому правило перевірки правильності висунутої гіпотези можна подати в наступному вигляді: гіпотеза  $c_1=c_2=\dots=c_n$  приймається, якщо  $F_p > F_T$ , і відкидається, якщо  $F_p > F_T$ ,

Таблиця 2

| Компонента дисперсії | Середній квадрат                                                        | Ступінь свободи |
|----------------------|-------------------------------------------------------------------------|-----------------|
| Між факторами        | $\frac{1}{N-1} \sum_{i=1}^N \sum_{l=1}^m (\bar{y}_i - \bar{y})^2$       | N-1             |
| Всередині серії      | $\frac{1}{Nm-m} \sum_{i=1}^N \sum_{l=1}^m (\bar{y}_{il} - \bar{y}_i)^2$ | N(m-1)          |
| Повна (загальна)     | $\frac{1}{Nm-1} \sum_{i=1}^N \sum_{l=1}^m (\bar{y}_i - \bar{y})^2$      | Nm-1            |

Таблиця 3

| Номер прикладу | Результати вимірювання |         |         |         |         |
|----------------|------------------------|---------|---------|---------|---------|
|                | 1                      | 2       | 3       | 4       | 5       |
| 1              | 98/-4                  | 98/-2   | 79/-21  | 96/-4   | 96/-4   |
| 2              | 107/+7                 | 111/+11 | 130/+30 | 128/+28 | 127/+27 |
| 3              | 119/+19                | 102/+2  | 87/-13  | 91/-9   | 102/+2  |

Таким чином, схема однофакторного дисперсійного аналізу може бути представлена у вигляді табл. 2, 3.

## 2. Двофакторний дисперсійний

Двофакторний дисперсійний аналіз застосовують для пов'язаних нормально розподілених вибірок. Дані подають у вигляді таблиці, у стовпчиках якої наводять дані, що відповідають певному рівню першого фактора, а в рядках – дані, що відповідають рівням другого. Таблиця даних має розмірність  $n \times k$ , де  $n$  і  $k$  – кількість рівнів першого та другого факторів, відповідно (табл. 4).

Таблиця 4

| Рівні фактора А | Рівні фактора В |           |     |           |
|-----------------|-----------------|-----------|-----|-----------|
|                 | 1               | 2         | ... | k         |
| 1               | $x_{11}$        | $x_{12}$  | ... | $x_{1k}$  |
| 2               | $x_{21}$        | $x_{22}$  | ... | $x_{2k}$  |
|                 |                 |           | ... |           |
| $p_i$           | $x_{ni1}$       | $x_{ni2}$ | ... | $x_{nik}$ |

Основною відмінністю від таблиці однофакторного дисперсійного аналізу є можлива неоднорідність даних у стовпцях, якщо вплив другого фактора є суттєвим. На практиці часто використовують і складніші таблиці двофакторного дисперсійного аналізу, зокрема такі, у яких кожна комірка містить набір даних (повторні вимірювання), що відповідають фіксованим значенням рівнів обох факторів.

Якщо припущення, необхідні для застосування двофакторного дисперсійного аналізу, не виконуються, то використовують непараметричний ранговий критерій Фрідмана (Фрідмана, Кендалла та Сміта), розроблений американським економістом Мілтоном Фрідманом наприкінці 1930 р.

Q-критерій Кокрена, запропонований В. Кочреном в 1937 р., використовують у випадках, коли групи однорідних суб'єктів піддаються впливам, кількість яких перевищує два, і для яких можливі два варіанти відгуків – умовно-негативний (0) та умовно-позитивний (1). Нульова гіпотеза полягає в рівності ефектів впливу.

Двофакторний дисперсійний аналіз дає можливість визначити існування ефектів обробки, проте не дає змоги встановити, для яких саме стовпців існує цей ефект. При вирішенні цієї проблеми застосовують метод множинних порівнянь Шеффе для пов'язаних вибірок [3].

### 3. Приклад виконання рангового однофакторного аналізу

Нехай ми маємо чотири вибірки, сформовані за допомогою пакета аналізу MS Excel як суміш 100 елементів нормально розподіленої вибірки з параметрами  $x = 20$ ,  $s = 3$  та рівномірно розподілених вибірок обсягом по 100 елементів кожна заданих, відповідно, на відрізках: [17; 22], [18; 22], [18; 22] та [17; 23]. В електронних таблицях MS Excel немає вбудованих засобів для реалізації рангового однофакторного аналізу Краскела - Уолліса. Але

його неважко здійснити за допомогою наявних функцій. Спочатку необхідно перетворити таблицю вихідних даних у таблицю значень рангів. Для цього можна використати функцію РАНГ (). Вікно задання її параметрів показано на рис. 1.

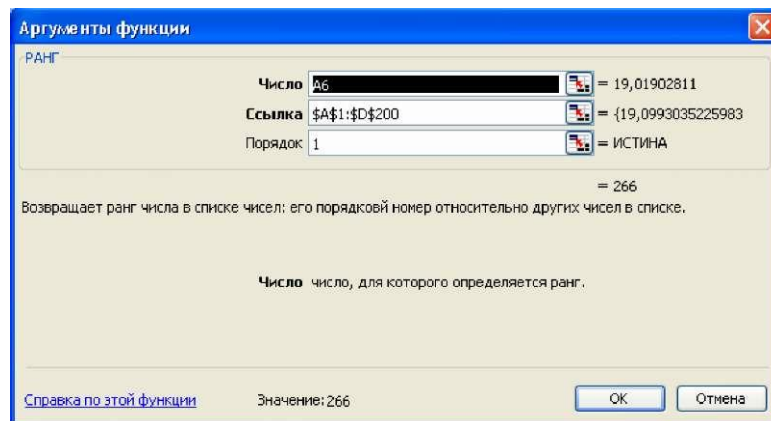


Рис. 1. Вікно задання параметрів функції РАНГ ()

У комірці "Число" вказуємо для якого саме значення таблиці вихідних даних необхідно обчислити ранг. У комірці "Ссылка" даємо посилання на весь діапазон, що містить вихідні значення (це посилання має бути абсолютним).

У комірці "Порядок" зазначаємо порядок ранжирування: 0 – за убаванням, інше число – за зростанням. Після цього розраховуємо значення критерію. У нашому випадку воно дорівнює 7,22. Як критичне візьмемо значення оберненої функції розподілу  $\chi^2$ , яке можна визначити за допомогою функції = ХИ2ОБР (0,05;3) = 7,82. Її аргументами є рівень значущості (0,05) та кількість степенів вільності (у нашому випадку 3). Бачимо, що розрахункове значення критерію дещо менше, ніж критичне. Тому немає підстав для відхилення нульової гіпотези про однорідність досліджуваних вибірок.

У пакеті SPSS реалізовано можливість перевірки однорідності вибірок числових даних, розподіл яких відрізняється від нормального закону, а також порядкових даних за допомогою рангового однофакторного аналізу Краскела-Уолліса.

Для реалізації цієї процедури необхідно увійти до діалогового вікна Analyze/Nonparametric Tests/K Independent Samples й обрати в ньому тест Краскела-Уолліса. Додатково ми можемо здійснити перевірку за медіанним тестом і критерієм Джонкхієра-Терпстри.

Результати наведено на рис. 2-6. Бачимо, що за результатами всіх тестів рівень значущості, що відповідає розрахунковому значенню критерію, є вищим, ніж 0,05. Тому ми можемо прийняти нульову гіпотезу про однорідність досліджуваних вибірок на рівні значущості 0,05.

### Ranks

| VAR00002 | N   | Mean Rank |
|----------|-----|-----------|
| VAR00001 | 200 | 370,98    |
| 1,00     |     |           |
| 2,00     | 200 | 394,42    |
| 3,00     | 200 | 432,20    |
| 4,00     | 200 | 404,40    |
| Total    | 800 |           |

### Test Statistics

|               | VAR00001 |
|---------------|----------|
| Chi-Square df | 7,221    |
| Asymp. Sig.   | 3,065    |

- a. Kruskal Wallis Test  
b. Grouping Variable: VAR00002

Рис. 2. Результати перевірки однорідності вибірок за критерієм Краскела-Уолліса

### Frequencies

|                   | VAR00002 |      |      |      |
|-------------------|----------|------|------|------|
|                   | 1,00     | 2,00 | 3,00 | 4,00 |
| VAR00001 > Median | 09       | 95   | 110  | 106  |
| <= Median         | 111      | 105  | 90   | 94   |

### Test Statistics

|             | VAR00001           |
|-------------|--------------------|
| N           | 300                |
| Median      | 20,0603            |
| Chi-Square  | 5,640 <sup>a</sup> |
| df          | 3                  |
| Asymp. Sig. | .131               |

- a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 100,0  
b. Grouping Variable: VAR00002

Рис. 3. Результати перевірки однорідності вибірок за медіанним тестом

### Jonckheere-Terpstra Test

|                                 |           |
|---------------------------------|-----------|
|                                 | VAR00001  |
| Number of Levels in VAR00002    | 4         |
| N                               | BOO       |
| Observed J-T Statistic          | 127124,00 |
| Mean J-T Statistic              | 120000,00 |
| Std. Deviation of J-T Statistic | 3654,221  |
| Std. J-T Statistic              | 1,950     |
| Asymp. Sig. (2-tailed)          | .051      |

a. Grouping Variable: VAR00002

Рис. 4. Результати перевірки однорідності вибірок за критерієм Джонкхієра-Терпстрі

Контрольні запитання.

1. Які завдання вирішують за допомогою дисперсійного аналізу?
2. Що називають факторами й відгуками у дисперсійному аналізі? Наведіть приклади.
3. Що називають рівнем фактора? Наведіть приклади.
4. Як визначають кількість і розміри інтервалів в однофакторному дисперсійному аналізі?
5. Які типи групування використовують у багатфакторному дисперсійному аналізі?
6. Якими є основні умови застосування однофакторного дисперсійного аналізу?
7. Доведіть основну тотожність дисперсійного аналізу.
8. Що являє собою факторна варіація та яку властивість даних вона характеризує?
9. Що являє собою залишкова варіація та яку властивість даних вона характеризує?
10. Які властивості даних характеризують оцінки дисперсії похибок, що використовуються у дисперсійному аналізі?
11. Які завдання вирішують за допомогою рангового однофакторного аналізу Краскела-Уолліса?

12. За яких умов можна використовувати ранговий однофакторний аналіз Краскела-Уолліса?
13. Який критерій є аналогом рангового однофакторного аналізу Краскела-Уолліса при порівнянні двох вибірок?
14. Які завдання вирішують за допомогою критерію Джонкхієра? За яких умов його доцільно використовувати?
15. Які завдання вирішують за допомогою критерію Бартлетта? За яких умов його можна використовувати?
16. Які завдання вирішують за допомогою G-критерію Кокрена? За яких умов його можна використовувати?
17. Яким є загальний вигляд адитивної моделі кількісного впливу досліджуваного фактора на відгук?
18. Що називають зсувом у дисперсійному аналізі? Як можна оцінити зсув кількісно?
19. Що називають лінійним контрастом в адитивній моделі? Наведіть приклади лінійних контрастів.
20. Які фактори визначають дисперсію лінійного контрасту?
21. Які завдання вирішують за допомогою методу множинних порівнянь Шеффе?
22. Які завдання вирішують за допомогою двофакторного дисперсійного аналізу? За яких умов можна використовувати цей метод?
23. Які властивості даних перевіряють за допомогою рангового критерію Фрідмана? За яких умов можна використовувати цей критерій?
24. Які властивості даних перевіряють за допомогою критерію Пейджа? За яких умов можна використовувати цей критерій?
25. Які властивості даних перевіряють за допомогою O-критерію Кокрена? За яких умов можна використовувати цей критерій?

## СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Цибульник С.О., Павловський О.М. Сучасні методи обробки інформації. Лекції [Електронний ресурс]: навч. посіб. для здобувачів ступеня доктора філософії спеціальності 151 «Автоматизація та комп'ютерно- інтегровані технології» / С.О. Цибульник, О.М. Павловський; КПІ ім. Ігоря Сікорського. Київ: КПІ ім. Ігоря Сікорського, 2021. 111 с.
2. Матеріали VI Міжнародної науково-технічної конференції молодих учених. Актуальні задачі сучасних технологій – Тернопіль 16-17 листопада 2017. 141 с.
3. Бахрушин В.Є. Методи аналізу даних : навчальний посібник. Запоріжжя : КПУ, 2011. 268 с.
4. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. К.: ІСЗІ КПІ ім. Ігоря Сікорського», 2018. 300 с.

Рекомендовано до друку рішенням вченої ради навчально-наукового інституту енергетики, автоматики і енергозбереження Національного університету біоресурсів і природокористування України (протокол № 11 від 17 грудня 2021 р.)

**НАВЧАЛЬНЕ ВИДАННЯ**  
**«Обробка інформації в комп'ютерно-інтегрованих системах автоматизації»**

**КОНСПЕКТ ЛЕКЦІЙ**  
з дисципліни  
“Обробка інформації в комп'ютерно-інтегрованих системах автоматизації”  
для аспірантів третього освітньо-наукового рівня вищої освіти  
зі спеціальності 151 – «Автоматизація та комп'ютерно-інтегровані технології»

Укладачі: докт. техн. наук, професор В.В. Коваль,  
канд. техн. наук, доцент М.О. Кіктєв.

Технічний редактор В.М. Поліщук  
Відповідальний за випуск С.В. Драгнєв  
Комп'ютерна верстка С.В. Драгнєв

Підписано до друку 17.12.21  
Ум. друк. арк. 6,9  
Наклад 20 прим.

Формат 60x84/16  
Обл. вид. арк. 6,2  
Зам. № 2919

Видавничий центр НУБіП  
03041, вул. Героїв Оборони, 15