**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**
**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ**
**«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Кафедра _____соціології і публічного управління_____
(назва кафедри, яка забезпечує викладання дисципліни)

# КОМПЛЕКС НАВЧАЛЬНО-МЕТОДИЧНОГО ЗАБЕЗПЕЧЕННЯ ДИСЦИПЛІНИ

# МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ

(назва навчальної дисципліни)

рівень вищої_____освіти другий (магістерський _____
перший (бакалаврський) / другий (магістерський)

галузь знань _____05 Соціальні й поведінкові науки _____
(шифр і назва)

спеціальність _____054 Соціологія_____
(шифр і назва)

освітня програма __Соціологічне забезпечення економічної діяльностіі____
(назви освітніх програм спеціальностей)

вид дисципліни ____/ професійна підготовка, обов'язкова_____
(загальна підготовка / професійна підготовка; обов'язкова/вибіркова)

форма навчання _____денна_____
(денна / заочна)

Харків – 2024 рік

**Силабус освітнього компонента**
Програма навчальної дисципліни

# МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ

***Шифр та назва спеціальності***
054 – Соціологія

***Інститут***
ННІ Соціально-гуманітарних технологій

***Освітня програма***
Соціологічне забезпечення економічної діяльності

***Кафедра***
Соціології і публічного управління (305)

***Рівень освіти***
Магістр

***Тип дисципліни***
Спеціальна (фахова), Обов'язкова

***Семестр***
2

***Мова викладання***
Українська, англійська

---

## Викладачі, розробники

**Бірюкова Марина Василівна**
Maryna.Biriukova@khpi.edu.ua
Доктор соціологічних наук, професор, доцент кафедри соціології і публічного управління
Автор 120 наукових та науково-методичних праць, у тому числі трьох одноосібних монографій та підручників.
Лектор з дисциплін: «Математичні методи в соціології», «Практикум з аналізу соціологічних даних», «Комп’ютерні технології організації соціологічних дисциплін», «Технології соціального проектування», «Методи багатомірного аналізу соціологічних даних».
Досвід роботи – 33 роки
Детальніше про викладача на сайті кафедри
https://web.kpi.kharkov.ua/sp/profesors-ko-vikladats-kij-sklad/

## Загальна інформація

### Анотація
Основними завданнями курсу є: вивчення методів наукових досліджень з теорії організації вибіркових спостережень, обробки та аналізу отриманої інформації, застосування багатовимірних методів та bigdata для соціального аналізу, ідентифікації та розпізнавання образів; моделювання і прогнозування соціальних процесів; використання інформаційних технологій для статистичного обгрунтування прийняття рішень при соціологічному забезпеченні економічної діяльності.

### Мета та цілідисципліни
Освоєння методологічних і методичних основ використовування методів багатовимірного аналізу та bigdata для дослідження природи соціальних явищ, для побудови багатовимірних моделей існування та функціонування соціальних об'єктів.

### Формат занять
Лекції, лабораторні роботи, самостійна робота, консультації. Підсумковий контроль – іспит.

### Компетентності
ЗК05. Здатність оцінювати та забезпечувати якість виконуваних робіт.

СК02. Здатність виявляти, діагностувати та інтерпретувати соціальні проблеми українського суспільства та світової спільноти.

СК03. Здатність проектувати і виконувати соціологічні дослідження, розробляти й обґрунтовувати їхню методологію.

СК04. Здатність збирати та аналізувати емпіричні дані з використанням сучасних методів соціологічних досліджень та цифрових технології.

СК07. Здатність розробляти та оцінювати соціальні проекти і програми.

### Результати навчання
ПР01. Аналізувати соціальні явища і процеси, використовуючи емпіричні дані та сучасні концепції і теорії соціології.

ПР02. Здійснювати діагностику та інтерпретацію соціальних проблем українського суспільства та світової спільноти, причини їхнього виникнення та наслідки.

ПР03. Розробляти і реалізовувати соціальні та міждисциплінарні проекти з урахуванням соціальних, економічних, правових, екологічних та інших аспектів суспільного життя.

ПР04. Застосовувати наукові знання, соціологічні та статистичні методи, цифрові технології, спеціалізоване програмне забезпечення для розв'язування складних задач соціології та суміжних галузей знань.

ПР05. Здійснювати пошук, аналізувати та оцінювати необхідну інформацію в науковій літературі, банках даних та інших джерелах.

ПР09. Планувати і виконувати наукові дослідження у сфері соціології, аналізувати результати, обґрунтовувати висновки.

## Обсяг дисципліни

Загальний обсяг дисципліни 180 год. (6 кредитівECTS): лекції – 32 год., семінарські заняття – 32 год., самостійна робота – 116 год.

## Передумови вивчення дисципліни (пререквізити)

Для успішного проходження курсу необхідно мати знання та практичні навички з наступних дисциплін: «Математичні методи в соціології", "Практикум з комп'ютерної обробки соціологічних даних", "Соціологічний супровід економічної діяльності". "Інтернет-дослідження економічної діяльності".

## Особливості дисципліни, методи та технології навчання

Під час проведення практичних занять з навчальної дисципліни передбачено пояснення алгоритму виконання практичних завдань та їх відпрацювання. Застосовуються наступні методи навчання: пояснювально-ілюстративний; репродуктивний(відпрацювання певних алгоритмів аналізу даних); частково-пошуковий або евристичний метод(під час виконання індивідуальних завдань). На практичних заняттях використовується проєктний підхід до навчання, гейміфікація, акцентується увага на застосуванні інформаційних технологій в організації соціологічних досліджень: проектна і командна робота, peer-to-peer, , кейси.

## Програма навчальної дисципліни

## Теми лекційних занять

### *Тема 1. Основні елементи формалізму*

Аналіз соціологічної інформації, зібраної в ході емпіричних соціологічних досліджень, є не просто сукупністю технічних прийомів і методів. Неодновимірність багатьох досліджуваних соціологом понять. Непрямий її прояв – порушення транзитивності відношення порядку. Метричне та неметричне БШ. Відповідні функції стресу. Неявне порівняння відстаней між близькістю, закладене у формулі функції стресу для метричного шкалірування. Поняття монотонної регресії, що використовується при розрахунку функції стресу для неметричного шкалірування. Важливість для соціології неметричного шкалірування. Формальні аспекти проблем розмірності шуканого евклідового простору і обертання, що визначають його осей координат.

### *Тема 2. Багатовимірне розгортання та індивідуальне багатовимірне шкалірування*

Постановка завдання; важливість врахування специфіки метрик окремих респондентів. Спосіб обліку таких метрик в індивідуальному БШ. Вид вхідних і вихідних даних, функції стресу в індивідуальному БШ. Одномірне розгортання. Обґрунтування необхідності переходу до простору довільної розмірності для успішного виконання завдання шкалірування. Модель ідеальної точки в багатовимірному випадку. Неметричне багатовимірне розгортання. Вид вихідних даних. Функція стресу. Специфіка вихідних даних (наявність двох видів точок, що відповідають об'єктам і респондентам відповідно). Особливості інтерпретації результатів.

### *Тема 3. Проблеми формування вихідних даних і інтерпретації результатів у багатовимірному шкалірування.*

Роль соціолога при отриманні даних, вихідних для багатовимірного шкалірування, та інтерпретації його результатів. Можливі способи одержання вихідних даних. Безпосереднє отримання близькості від респондентів, класифікація відповідних способів опитування; проблеми, що постають при такому способі збору даних. Приклади розрахунку матриці близькості на основі аналізу достатньо надійних даних іншого роду. Робота з БШ статистичними програмами - процедура БШ доступна в більшості статистичних програм. Існує вибір між метричним БШ (який дозволяє працювати з інтервалами чи даними про співвідношення рівня), і неметричним БШ (який працює з порядковими даними). Використання формальних та неформальних методів при інтерпретації результатів багатовимірного шкалірування. Значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей.

### *Тема 4. Канонічний аналіз. Загальне уявлення про методи, які засновані на моделях частот*

Загальне уявлення про моделювання частот таблиці спряженості. Змістовне розуміння таких моделей, їх роль для соціолога. Мультиплікативні та адитивні моделі частот. Роль логарифмування мультиплікативної моделі. Можливість різного розуміння як сенсу розглянутих вкладів, так і того "середнього" рівня, з яким порівнюються спостерігаються частоти в процесі їх моделювання. Канонічний кореляційний аналіз – один із мето- дів багатовимірного аналізу даних. Необхідність сполучення моделі, закладеної в конкретному методі оцифровки, з вмістом розглянутої задачі. Приклад моделі такого роду – модель, використовувана в методі шкалірування, званому методом послідовних розбивок. Канонічний аналіз як метод оцифровки і метод вимірювання зв'язку між двома номінальними ознаками зі "спільними альтернативами". Моделі частот, що відповідають канонічному аналізу. Побудова соціологічних індексів за допомогою техніки канонічного аналізу. Вирішення проблеми зважування складових індекс ознак.

### *Тема 5. Логлінейний аналіз*

Логлінейний аналіз - метод багатовимірного статистичного аналізу для вивчення таблиць спряженості. Логлінейний аналіз дозволяє статистично перевіряти гіпотезу про систему одночасно мають місце парних і множинних взаємозв'язків в групі ознак, виміряних за номінальними шкалами. Багатовимірний статистичний аналіз. Моделі частот, що відповідають логлінейному аналізу. Насичена модель. Мета переходу до логарифмів частот. Сенс вкладів різної розмірності. Різне розуміння залежної ознаки: кількісна ознака в дисперсійному аналізі, кількісна або номінальна – в номінальному регресійному і частота, що стоїть в клітці багатовимірної таблиці спряженості,  – в логлінейном аналізі. Різні можливості пошуку поєднань значень предикторів: перевірка гіпотез про наявність багатовимірних зв'язків у логлінейном аналізі і можливість пошуку найбільш дієвих поєднань в методі послідовних розбивок і регресійному аналізі, заздалегідь заданий набір поєднань значень предикторів в дисперсійному аналізі.

### *Тема 6. Причинний аналіз. Стратегія аналізу структури взаємозв'язків ознак*

Поняття причини в соціології. Принципова неможливість повністю його формалізувати. Роль статистичних методів при вивченні причинних відносин. Граф причинних зв'язків. Структурні коефіцієнти. Вхідні (зовнішні, незалежні) і вихідні (внутрішні, залежні) змінні. Правила редукції причинних схем та формування рівнянь. Повторення принципів побудови часткових коефіцієнтів кореляції і регресії. Важливість для соціолога вивчення відповідних зв'язків. Різниця між статистичним та причинним зв'язком. Поняття "помилкової" кореляції. Основні причинні схеми, що призводять до їх появи. Проблема формалізації завдання вивчення причинно-наслідкових відносин в соціології.

Поняття структури багатовимірної випадкової величини. Формування узагальнених показників на базі аналізу структури зв'язків ознак. Комплексне використання декількох методів вивчення зв'язків між ознаками для вирішення соціологічних задач (аналіз структури випадкової величини; факторний і дисперсійний аналіз; пошук детермінуючих поєднань значень предикторів).

### *Тема 7. Завдання розпізнавання образів. Поняття автоматичної класифікації об'єктів*

Класифікація як один із фундаментальних процесів у науці. Ознаковий простір. Задача класифікації як пошук згущення точок – моделей об'єктів в ознаковому просторі.

Виділення задачі автоматичної класифікації об'єктів (синоніми: багатовимірна класифікація, розпізнавання образів без вчителя, кластерний аналіз, таксономія).

Класифікація як один із фундаментальних процесів у науці. Ознаковий простір. Задача класифікації як пошук згущення точок – моделей об'єктів в ознаковому просторі.Виділення задачі автоматичної класифікації об'єктів (синоніми: багатовимірна класифікація, розпізнавання образів без вчителя, кластерний аналіз, таксономія).

### *Тема 8. Проблема "стикування" змісту і формалізму при використанні алгоритмів класифікації*

Специфіка рішення соціологічних завдань побудови типології за допомогою методів автоматичної класифікації. Сенс протиставлення термінів "класифікація" і "типологія". Підстава типології. Роль апріорних уявлень дослідника про шуканих типах у виборі і реалізації алгоритму, інтерпретації результатів його застосування. Виділення основних формальних елементів алгоритмів автоматичної класифікації, що вимагають стикування зі змістовними концепціями соціолога.

### *Тема 9. Функції відстані між об'єктами*

Аксіоматичне визначення функції відстані і ролі цієї функції в соціології. Приклади непридатності евклідової відстані з точки зору апріорного змістовного розуміння типів об'єктів.

Можливість використання евклідової відстані в розглянутих прикладах за рахунок зміни ознакового простору. Сучасний аналіз даних обумовлюється способами отримання величин, методами їх обробки й залежить від розвитку математичних методів і моделювання. Функції відстані, відмінні від евклідова: зважене евклидово, сіті-блок, Махаланобіса, Хеммінгово.

### *Тема 10. Основні види процедур класифікації. Відстані між класами*

Актуальність дослідження сутності та методів багатовимірного аналізу соціологічної інформації обумовлена специфікою соціальної реальності, що завжди уявляється як складний, багатогранний та багатозначний феномен, який інтегрує багатовимірність суспільства з багатовимірністю внутрішнього світу окремої людини. Виділення ієрархічних і неієрархічні алгоритмів класифікації. Багатовимірний статистичний аналіз (у широкому значенні) - розділ математичної статистики, що поєднує методи вивчення даних, які характеризують багатовимірні об'єкти. Агломеративні та дівізімні алгоритми. Причини необхідності розгляду відстаней між класами в ієрархічних процедурах. Алгоритм найближчого сусіда як приклад способу класифікації, що використовує такі відстані.

### *Тема 11. Гіпотези про розташування об'єктів у ознаковому просторі*

Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації. Обумовленість цих гіпотез апріорними уявленнями дослідника про типи об'єктів. Основні види гіпотез: компактності, зв'язності (безперервності), унімодального розподілу. Факторний аналіз найбільш яскраво відображує риси багатомірного аналізу в частині дослідження зв'язку між ознаками. Кластерний аналіз ці риси відображує з боку класифікації об'єктів. Загальне уявлення про розмиті класифікації. Роль функції належності у відповідних алгоритмах. Доцільність комплексного використання декількох алгоритмів класифікації в соціологічних завданнях побудови типології. Змістовні уявлення соціолога про типи та умови вибору кроку розбиття при інтерпретації результатів. Коригування результатів класифікації з метою забезпечення відповідності класифікації і типології.

### *Тема 12. Поняття інтерпретації вихідних даних і основні методологічні принципи використання методів аналізу даних в соціології*

Інтерпретація вихідних даних як одне з основних ланок "стикування" соціології і математики. Основні фактори, що визначають інтерпретацію вихідних даних: апріорні уявлення дослідника про спосіб породження цих даних (у тому числі – про моделі сприйняття респондентами пропонованих ним питань, об'єктів, про ймовірнісну природу даних і т. д.); мета дослідження; концептуальні уявлення соціолога про досліджуване явище; характер моделі явища, "закладеної" в математичному методі, використання якого планується; розгляд спостережуваних змінних як непрямих показників латентних факторів, насправді цікавлять дослідника і т. п.

Виділення методологічних принципів, дотримання яких є необхідним для того, щоб аналіз соціологічних даних був ефективний, не відводив соціолога в

сторону від реальності: забезпечення певної однорідності вихідних даних; облік моделі, "закладеної" в кожному методі аналізу даних, при виборі алгоритму аналізу, два основні принципи інтерпретації результатів аналізу: необхідність її узгодження з інтерпретацією вихідних даних і заповнення при її здійсненні тих втрат, які мали місце при переході до формалізму; необхідність комплексного використання декількох методів для вирішення одного завдання і т. п.

## *Тема 13. Дані. Метадані*

Згідно з ГОСТ, дані – подання інформації у формалізованому вигляді, придатному для передачі, інтерпретації та обробки.

Вихідне поняття даних - філософське, воно виникає в епістемології під час розгляду основою проблеми гносеології – пізнаваності світу, пошуку та осмислення істини. Процедури верифікації чи фальсифікації даних створюють інформацію, осмислення істини створює знання.

Життєвий цикл даних – це послідовність етапів, яку конкретна порція даних проходить від початкового етапу створення чи отримання до моменту архівації чи видалення.

При зборі даних виникають метадані, що містять будь-яку інформацію про зібрані дані.

Огляд основних аналітичних інструментів роботи з Bigdata соціальних науках (Python, R, SAS, та ін). Читання та запис даних, формати файлів. Завантаження даних із різних джерел. Взаємодія з базами даних. Читання даних із Excel. Робота з CSV файлами та даними у форматі JSON. Парсинг простих даних XML. Читання даних із таблиць HTNL. Читання даних із файлу SAS. Взаємодія з HTML та Web API.

## *Тема 14. Великі дані. Системи керування великими даними*

Великі дані можуть бути різних типів. Інформацію, отриману в результаті обліку або вимірювання будь-яких об'єктів або параметрів, називають майстер-даними (MasterData). Наприклад, облік кількості, виміри координаті швидкостей конкретних молекул - це майстер-дані.

Транзакційні дані (в англомовній літературі застосовуються терміни TransactionalData, ApplicationSpecificData, OperationalData) – це дані, що відображають результат виконання будь-яких операцій. Транзакційні дані описують взаємодію об'єктів один з одним або з навколишнім світом, які можна отримати за допомогою обробки майстер-даних.

Ретроспективні дані (Historicaldata) – це дані, забезпечені позначки часу.

Посилальні дані (довідники, НСІ, нормативно-посилальна інформація,

ReferenceData, LookupData, Dictionaries) – це базові незмінні дані, заздалегідь відомі із зовнішніх джерел, такі як нормативи, скорочення, акроніми, словники, стандарти.

Формат даних. Структуровані дані мають заздалегідь визначений формат. Напівструктуровані або слабо структуровані дані -це дані, які часто зібрані з різних джерел.

### *Тема 15. Програмні платформи та системи для Великих даних*

В даний час використовується значна кількість платформ та систем Великих даних. Системи обробки великих даних єфреймворками, тобто каркасами, для використання яких необхідно з'єднати їх з іншими фреймворками, прикладним програмним забезпеченням користувача та системою зберігання даних.

В аналітичному звіті BigDataAnalyticsMarketStudy, 2017 Edition наводиться така діаграма інфраструктур Великих даних, впроваджених на підприємствах, представлена у розрізі розмірів підприємств

Розподілена обробка даних тісно пов’язана з паралельною обробкою даних. Однак така обробка завжди виконується за допомогою окремих машин у кластері, підключеному до мережі. Розподілена обробка даних - це метод виконання прикладних програм групою систем. Користувач може працювати з мережевими службами та прикладними процесами, розташованими в кількох взаємопов’язаних абонентських системах. Розподілена обробка даних підвищує ефективність інформаційних потреб користувачів і забезпечує ефективність та результативність рішень.

### *Тема 16. Машинне навчання за допомогою бібліотеки Scikit-learn.*

Види машинного навчання. Основні бібліотеки машинного навчання Python (Scikit-learn, Keras, TensorFlow). Створення тренувальних наборів -передобробка даних. Точність та достовірність моделі. Вибір найкращої моделі.

Кроки типового практичного сценарію машинного навчання. Завантаження набору даних. Дослідження даних за допомогою Pandas. Візуалізація ознак за допомогою Matplotlib. Розбиття даних для навчання та тестування. Створення моделі. Вивчення моделі. Тестування моделі.

Налаштування параметрів моделі та оцінка її точності. Формування прогнозів на підставі «живих» даних, які ще невідомі моделі.

Функціонал бібліотеки Scikit-Learn. Класифікація за допомогою К-сусідів.

Лінійні моделі для регресії та класифікації (модель лінійної регресії,логістична регресія, та ін). Наївні байєсівські класифікатори. Дерева рішень та випадковий ліс. Спосіб опорних векторів. Основи нейронних мереж.

Метод основних компонентів. Алгоритми кластеризації (кластеризація методом К-середніх, ієрархічна кластеризація, та ін).

## Теми практичних занять

### *Тема 1. Основні елементи формалізму*

Проблеми неодновимірності багатьох досліджуваних соціологом понять. Особливості вивчення простору сприйняття соціологічних явищ та процесів – основне завдання БШ. Ідеї Кумбса щодо урахування можливості упорядкування відстаней між об'єктами. Векторна модель або модель ідеальної крапки як основа БШ. Функція відстані (аксіоматичне визначення). Відповідні функції стресу. Простір сприйняття респондентами запропонованих їм об'єктів. Формальне визначення близькості. Вихідні дані для БШ – матриця близькості між об'єктами. Метричне та неметричне БШ. Формальні аспекти проблем розмірності шуканого евклідового простору і обертання, що визначають його осей координат. Розв’язання практичних завдань.

### *Тема 2. Багатовимірне розгортання та індивідуальне багатовимірне шкалірування*

Постановка завдання важливість врахування специфіки метрик окремих респондентів. Вид вхідних і вихідних даних, функції стресу в індивідуальному БШ. Одномірне розгортання. Обґрунтування необхідності переходу до простору довільної розмірності для успішного виконання завдання шкалірування. Неметричне багатовимірне розгортання. Особливості інтерпретації результатів. Спосіб обліку таких метрик в індивідуальному БШ. Модель ідеальної точки в багатовимірному випадку. Функція стресу. Специфіка вихідних даних (наявність двох видів точок, що відповідають об'єктам і респондентам відповідно).Розв’язання практичних завдань.

### *Тема 3. Проблеми формування вихідних даних і інтерпретації результатів у багатовимірному шкалірування*

Роль соціолога при отриманні даних, вихідних для багатовимірного шкалірування та інтерпретації його результатів. Класифікація відповідних способів опитування; проблеми, що постають при такому способі збору даних. Приклади розрахунку матриці близькості на основі аналізу достатньо надійних даних іншого роду. Використання формальних та неформальних методів при інтерпретації результатів багатовимірного шкалірування. Значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей. Можливі способи одержання вихідних даних. Проблеми застосування статистичних методів в соціології. Основні функції та процедури аналізу даних. Значення змістовних концепцій дослідника при

вирішенні проблем вибору розмірності евклідова простору і повороту його осей. Створення багатовимірних таблиць за допомогою вторинних змінних. Загальна характеристика сучасних програмних засобів аналізу соціологічних даних. Розв’язання практичних завдань.

### *Тема 4. Канонічний аналіз. Загальне уявлення про методи, які засновані на моделях частот*

Загальне уявлення про моделювання частот таблиці спряженості. Мультиплікативні та адитивні моделі частот. Роль логарифмування мультиплікативної моделі. Основне завдання канонічного аналізу. Принципи їх отримання на основі аналізу таблиці спряженості. Моделі частот, що відповідають канонічному аналізу. Зв'язок канонічних коефіцієнтів кореляції з критерієм «хі-квадрат". Загальне уявлення про оцифрування значень номінальних ознак. Канонічний аналіз як метод оцифровки і метод вимірювання зв'язку між двома номінальними ознаками зі "спільними альтернативами". Поняття зв'язку між двома групами ознак. Послідовність канонічних коефіцієнтів кореляції.

Принципи отримання канонічних коефіцієнтів кореляції на основі аналізу таблиці спряженості.
Використання канонічної кореляції в аналізі таблиць спряженості.
Необхідність сполучення моделі, закладеної в конкретному методі оцифровки.
Побудова соціологічних індексів за допомогою техніки канонічного аналізу.
Вирішення проблеми зважування складових індекс ознак.
Розв’язання практичних завдань.

### *Тема 5. Логлінейний аналіз*

Причини відмінності реального розподілу від рівномірного. Моделі частот, що відповідають логлінейному аналізу. Насичена модель. Мета переходу до логарифмів частот. Гіпотези про взаємозв'язок ознак. Їх роль при побудові моделей частот. Розрахунок коефіцієнтів логлінейной моделі для двовимірного випадку. Відносини переважання. Інтерпретація коефіцієнтів через відносини переважання (для моделі довільної розмірності). Порівняння логлінейного аналізу з номінальним регресійним і дисперсійним аналізом, а також з методом послідовних розбивок. Порівняння здійснюється на змістовному рівні. Різне розуміння залежної ознаки: кількісна ознака в дисперсійному аналізі, кількісна або номінальна – в номінальному регресійному і частота, що стоїть в клітці багатовимірної таблиці спряженості, – в логлінейном аналізі. Неможливість отримання нового знання на основі аналізу рівномірного розподілу (суть аналізу даних – вивчення змін, порівняння показників різного роду). Сенс вкладів різної розмірності. Роль критерію "хі-квадрат" при використанні логлінейного аналізу. Відносини переважання. Інтерпретація коефіцієнтів через відносини

переважання (для моделі довільної розмірності). Різні можливості пошуку поєднань значень предикторів: перевірка гіпотез про наявність багатовимірних зв'язків у логлінейном аналізі і можливість пошуку найбільш дієвих поєднань в методі послідовних розбивок і регресійному аналізі, заздалегідь заданий набір поєднань значень предикторів в дисперсійному аналізі. Розв’язання практичних завдань.

### *Тема 6. Причинний аналіз. Стратегія аналізу структури взаємозв'язків ознак*

Граф причинних зв'язків. Повторення принципів побудови часткових коефіцієнтів кореляції і регресії. Важливість для соціолога вивчення відповідних зв'язків. Поняття "помилкової" кореляції. Основні причинні схеми, що призводять до їх появи. Обчислення ковариаций (кореляцій) між будь-якими двома ознаками на основі графа зв'язків. Структурні рівняння. Обчислення структурних коефіцієнтів. Їх зв'язок з частковими коефіцієнтами регресії. Основна теорема причинного аналізу. Її роль у вивченні статистичних залежностей. Поняття структури багатовимірної випадкової величини. Формування узагальнених показників на базі аналізу структури зв'язків ознак. Роль статистичних методів при вивченні причинних відносин. Структурні коефіцієнти. Вхідні (зовнішні, незалежні) і вихідні (внутрішні, залежні) змінні. Правила редукції причинних схем та формування рівнянь. Різниця між статистичним та причинним зв'язком. Вивчення статистичних зв'язків на основі причинних схем як основне завдання причинного аналізу. Поняття допоміжної теорії вимірювань Блейлока. Причинний аналіз як концептуальний підхід до вивчення соціальних явищ. Проблема формалізації завдання вивчення причинно-наслідкових відносин в соціології. Комплексне використання декількох методів вивчення зв'язків між ознаками для вирішення соціологічних задач (аналіз структури випадкової величини; факторний і дисперсійний аналіз; пошук детермінуючих поєднань значень предикторів). Розв’язання практичних завдань.

### *Тема 7. Завдання розпізнавання образів. Поняття автоматичної класифікації об'єктів*

Класифікація як один із фундаментальних процесів у науці. Загальне уявлення про завдання розпізнавання образів (синоніми: образ, клас, кластер, таксон; неоднозначність трактування термінів в літературі). Виділення завдань: пошук класів, опис класів, визначення найбільш ефективної системи ознак. Виділення задачі автоматичної класифікації об'єктів (синоніми: багатовимірна класифікація, розпізнавання образів без вчителя, кластерний аналіз, таксономія). Ознаковий простір. Задача класифікації як пошук згущення точок – моделей об'єктів в ознаковому просторі. Роль наявності або відсутності навчальної вибірки. Розв’язання практичних завдань.

### *Тема 8. Проблема "стикування" змісту і формалізму при використанні алгоритмів класифікації*

Специфіка рішення соціологічних завдань побудови типології за допомогою методів автоматичної класифікації. Сенс протиставлення термінів "класифікація" і "типологія". Виділення основних формальних елементів алгоритмів автоматичної класифікації, що вимагають стикування зі змістовними концепціями соціолога. Підстава типології. Роль апріорних уявлень дослідника про шуканих типах у виборі і реалізації алгоритму, інтерпретації результатів його застосування.

Розв’язання практичних завдань.

### *Тема 9. Функції відстані між об'єктами*

Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації. Основні види гіпотез: компактності, зв'язності (безперервності), унімодального розподілу. Приклади соціологічних завдань побудови типології, для яких була б розумна кожна гіпотеза. Приклади алгоритмів, що шукають закономірності розташування точок у ознаковому просторі, що відповідають кожній з гіпотез: алгоритм Форель (гіпотеза компактності), алгоритм найближчого сусіда (гіпотеза зв'язності), алгоритм, заснований на виділенні локальних максимумів функції приналежності (гіпотеза унімодального розподілу ). Роль функції належності у відповідних алгоритмах. Змістовні уявлення соціолога про типи та умови вибору кроку розбиття при інтерпретації результатів. Коригування результатів класифікації з метою забезпечення відповідності класифікації і типології. Розв’язання практичних завдань.

### *Тема 10. Основні види процедур класифікації. Відстані між класами*

Виділення ієрархічних і неієрархічні алгоритмів класифікації. Агломеративні та дівізімні алгоритми. Оптимізація розбиття в сенсі максимізації заздалегідь обраного функціоналу якості як один з основних елементів формалізму в неієрархічні алгоритмах класифікації. Основний змістовний сенс оптимізації. Сенс вимірювання близькості між класами в таких випадках. Способи вимірювання сумарних оцінок близькості один до одного об'єктів усередині класів. Розв’язання практичних завдань.

### *Тема 11. Гіпотези про розташування об'єктів у ознаковому просторі*

Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації. Приклади соціологічних завдань побудови типології, для яких була б розумна кожна гіпотеза. Загальне уявлення про розмиті класифікації. Роль функції належності у відповідних алгоритмах. Змістовні уявлення соціолога про

типи та умови вибору кроку розбиття при інтерпретації результатів. Розв’язання практичних завдань.

### *Тема 12. Поняття інтерпретації вихідних даних і основні методологічні принципи використання методів аналізу даних в соціології*

Інтерпретація вихідних даних як одне з основних ланок "стикування" соціології і математики. Виділення методологічних принципів, дотримання яких є необхідним для того, щоб аналіз соціологічних даних був ефективний, не відводив соціолога в сторону від реальності: забезпечення певної однорідності вихідних даних; облік моделі, "закладеної" в кожному методі аналізу даних, при виборі алгоритму аналізу, два основні принципи інтерпретації результатів аналізу: необхідність її узгодження з інтерпретацією вихідних даних і заповнення при її здійсненні тих втрат, які мали місце при переході до формалізму; необхідність комплексного використання декількох методів для вирішення одного завдання і т. д. Розв’язання практичних завдань.

### *Тема 13.Дані. Метадані*

Створення даних (DataGeneration/DataCapture). Обслуговування даних (DataMaintenance). Синтез даних (DataSynthesis). Використання даних (DataUsage). Публікація даних (DataPublication). Архівація даних (DataArchival). Знищення даних (DataPurging) Розв’язання практичних завдань.

### *Тема 14. Великі дані. Системи керування великими даними*

Розподілені файлові системи. Розподілені фреймворки. Бенчмаркінг. Серверне програмування. Планування. Системи розгортання. Розв’язання практичних завдань.

### *Тема 15. Програмні платформи та системи для Великих даних*

Системи керування потоками даних. Системи зберігання Великих даних. Платформи Великих даних. Обробка даних у реальному часі. Системи керування Великими даними. Аналітичні платформи. Розв’язання практичних завдань.

### *Тема 16.Машинне навчання за допомогою бібліотеки Scikit-learn.*

Кроки типового практичного сценарію машинного навчання. Завантаження набору даних. Дослідження даних за допомогою Pandas. Візуалізація ознак за допомогою Matplotlib. Налаштування параметрів моделі та оцінка її точності. Функціонал бібліотеки Scikit-Learn. Класифікація за допомогою К-сусідів.

Лінійні моделі для регресії та класифікації (модель лінійної регресії, логістична регресія, та ін). Дерева рішень та випадковий ліс. Основи нейронних мереж. Алгоритми кластеризації (кластеризація методом К-середніх, ієрархічна кластеризація, та ін). Розв'язання практичних завдань.

## Теми лабораторних робіт

Лабораторних занять не передбачено.

## Самостійна робота

Самостійна робота за курсом складається із самостійного вивчення студентами тем та питань, які не викладаються на заняттях, виконання індивідуальних завдань. Студентам також рекомендуються додаткові матеріали (відео, статті) для самостійного вивчення та аналізу.

## Література та навчальні матеріали

Основна література

1. Горбачик А.П., Сальнікова С.А. Аналіз даних соціологічних досліджень засобами SPSS: Навч. посіб.- Луцьк, 2008. – 164 с. IBM SPSS 20 інструкція користувача// https://www.xn--80aaexjatkpdggghih8b1a2yhv.com.ua/ibm/spss-20/%D1%96%D0%BD%D1%81%D1%82%D1%80%D1%83%D0%BA%D1%86%D1%96%D1%8F-%D0%BA%D0%BE%D1%80%D0%B8%D1%81%D1%82%D1%83%D0%B2%D0%B0%D1%87%D0%B0.

2. Паніотто В.І., Максименко В. С., Харченко Н.М. Статистичний аналізсоціологічних даних. - Київ, 2004. – 270 с. Литвин В.В. Аналіз даних та знань: підручник/ В.В. Литвин, В.В. Пасічник, Ю.В. Нікольський.- Л.: Магнолія, 2020.- 276с. (базовий підручник).

Допоміжна література

3. Лупан І.В., Авраменко О.В., АкбашК.С.Комп'ютерні статистичні пакети: навчально-методичнийпосібник. - 2-е вид. - Кіровоград: 'КОД". 2015. - 230 с. - http://dspace.cuspu.edu.ua/jspui/bitstream/123456789.

4. MakingSenseofMultivariateData Analysis//https://us.sagepub.com/en-us/nam/book/making-sense-multivariate-data-analysis

5. Бахрушин В.Є.Методи аналізу даних: навчальний посібник для студентів В.Є. Бахрушин. - Запоріжжя : КПУ, 2011. - 26В с. - http://web.kpi.kharkov.ua/auts/wp-content/uploads/sites/67/2017/02/DAMAP_Ivashko_posobie2.pdf

6. Інтелектуальний аналіз даних: практикум/ М.Т. Фісун, І.О. Кравець, П.П. Казмірчук.- Л.: Новий Світ-2000, 2020.- 162с. Гладун А.Я., Рогушина Ю. В. DataMining: пошук знань в даних. Київ. ТОВ «ВД «АДЕФ- Україна», 2016. — 452 с..

## Система оцінювання

### Критерії оцінювання успішності студента та розподіл балів

100% підсумкової оцінки складаються з результатів оцінювання у вигляді іспиту (20%) та поточного оцінювання (80%). Іспит: виконання разрахункового завдання та усна доповідь. Поточне оцінювання: 16 онлайн тестів за темами (48%), два індивідуальні завдання (22%)та два разрахункових завдання (10%)

### Шкала оцінювання

| *Сума балів* | *Національнаоцінка* | *ECTS* |
|---|---|---|
| 90–100 | Відмінно | A |
| 82–89 | Добре | B |
| 75–81 | Добре | C |
| 64–74 | Задовільно | D |
| 60–63 | Задовільно | E |
| 35–59 | Незадовільно (потрібне додаткове вивчення) | FX |
| 1–34 | Незадовільно (потрібне повторне вивчення) | F |

## Норми академічної етики і політика курсу

Студент повинен дотримуватися «Кодексу етики академічних взаємовідносин та доброчесності НТУ «ХПІ»: виявляти дисциплінованість, вихованість, доброзичливість, чесність, відповідальність. Конфліктні ситуації повинні відкрито обговорюватися в навчальних групах з викладачем, а при неможливості вирішення конфлікту – доводитися до відома співробітників дирекції інституту.

Нормативно-правове забезпечення впровадження принципів академічної доброчесності НТУ «ХПІ» розміщено на сайті: http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/

## Погодження

Силабус погоджено Дата погодження, підпис *Завідувач кафедри*

Володимир МОРОЗ

30.06.2023 

Дата погодження, підпис

***Гарант ОП***
Юрій КАЛАГІН

30.06.2023 

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

НАВЧАЛЬНО-МЕТОДИЧНІ МАТЕРІАЛИ ЛЕКЦІЙ

# МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ

**для студентів спеціальності 054 «Соціологія»**

**Харків − 2024**

## Тема 1. Основні елементи формалізму

1. Неодновимірність багатьох досліджуваних соціологом понять.
2. Простір сприйняття респондентами запропонованих їм об'єктів.
3. Вивчення простору сприйняття – основне завдання БШ.

Аналіз соціологічної інформації, зібраної в ході емпіричних соціологічних досліджень, є не просто сукупністю технічних прийомів і методів. Це ключовий етап усього дослідження, в якому відбувається конкретна перевірка відповідності зібраної інформації тим моделям соціальних явищ, які, явно чи приховано, є у соціолога. І більш того, в процесі аналізу визначають і перевіряються нові моделі, які відповідно відображають ті закономірності, які є в зібраних даних. Неодновимірність багатьох досліджуваних соціологом понять. Непрямий її прояв – порушення транзитивності відношення порядку.

Простір сприйняття респондентами запропонованих їм об'єктів. Його латентність. Вивчення простору сприйняття – основне завдання БШ. Інші завдання БШ (пониження розмірності досліджуваного ознакового простору, візуалізація даних). Їх роль в соціології.

Ідеї Кумбса щодо урахування можливості упорядкування відстаней між об'єктами, необхідність аналізу моделі сприйняття респондентом запропонованих йому об'єктів – векторної або моделі ідеальної крапки як основа БШ.

Формальне визначення близькості. Вихідні дані для БШ – матриця близькості між об'єктами. Функція відстані (аксіоматичне визначення). Евклідова відстань. Евклідова простір. Вихідна інформація – координати об'єктів, що шкаліруються в евклідовому просторі, матриця відстаней між ними. Вимога відповідності між структурами матриці близькості і матриці відстаней.

Метричне та неметричне БШ. Відповідні функції стресу. Неявне порівняння відстаней між близькістю, закладене у формулі функції стресу для метричного шкалірування. Поняття монотонної регресії, що використовується при розрахунку функції стресу для неметричного шкалірування.

Важливість для соціології неметричного шкалірування. Формальні аспекти проблем розмірності шуканого евклідового простору і обертання, що визначають його осей координат.

Цілі та завдання аналізу даних у соціології

Аналіз даних це поняття означає сукупність дій, здійснювані дослідником у процесі вивчення деяких даних з формування певного ставлення до характері описуваного явища.

Предметом є техніка формування певного уявлення.

Основна мета аналізу даних - виявлення (підтвердження, коригування) якихось дослідників, що цікавлять статистичних закономірностей; або, іншими словами, - певного роду стиск, усереднення інформації, що міститься в даних.

Цілі та завдання аналізу даних у соціології

Завдання пошуку закономірності іноді ототожнюють із завданням пояснення цікавого дослідника явища (нагадаємо, що головний сенс пояснення полягає у підведенні явища, що пояснюється, під який-небудь закон, явище – це не обов'язково наша змістовна закономірність).

Поряд з поясненням досліджуваного явища, доцільно завжди мати на увазі принаймні ще дві мети: опис вихідних даних і здійснюване на основі виявленої закономірності передбачення того чи іншого явища.

Опис - мета, досягти яку часто буває необхідно перш ніж безпосередньо приступати до пошуку основної дослідника, що цікавить закономірності.

Пророцтво теж часто вважається основною метою наукового дослідження і з цим важко сперечатися (афоризм О.Конта «Знати, щоб передбачити»).

Цілі та завдання аналізу даних у соціології

Потрібно згадати про розуміння як одну з пізнавальних функцій соціології з величезної важливості досягнення розуміння досліджуваного об'єкта (людини) у кожному соціологічному дослідженні. "Розуміння" зазвичай досягається за допомогою м'яких методів дослідження.

Основними завданнями, які вирішує аналіз даних, є:

1. Класифікація об'єктів – пошук однотипних груп об'єктів, створення типології.

2. Стиснення інформації:• Одномірний аналіз – описова статистика;• Багатомірний аналіз – зв'язок між ознаками;• Пошук латентних змінних.

Статистична закономірність у соціології

У науці прийнято виділяти дві основні форми закономірного зв'язку явищ, що відрізняються за характером передбачень, що випливають з них: динамічні і статистичні закономірності.

У законах динамічного типу передбачення має точний, певний однозначний вигляд; у статистичних законах передбачення носить не достовірний, лише ймовірнісний характер.

Нас цікавлять переважно статистичні закономірності, звані закономірності " у середньому " .

Статистична закономірність виникає як наслідок взаємодії великої кількості елементів, складових сукупність, і характеризують й не так поведінка окремого елемента сукупності, скільки всю сукупність загалом.

Статистична закономірність у соціології

Статистичні закономірності цілком адекватно описують масові явища випадкового характеру, саме такого роду явища і вивчає зазвичай соціолог.

Найчастіше, говорячи про статистичність соціальних закономірностей, дослідники мають на увазі закони розвитку великих соціальних груп та суспільства загалом. При цьому подібна статистичність зазвичай розглядається в контексті аналізу відомої дилеми про співвідношення загальних закономірностей розвитку суспільства та свободою волі окремої людини.

Проблема унікального та середнього.

Знаходження різноманітних статистичних закономірностей є звичною справою кожного соціолога, який проводить емпіричне дослідження. Але нам видається некоректним, коли статистичний підхід пов'язується лише з великими групами чи суспільством загалом.

Статистичні моделі можуть використовуватися і при спробі "зрозуміти" окрему людину, і при вивченні різного роду груп людей, у тому числі суспільства в цілому.

Адекватно відбивають суть статистичного підходу щодо окремої людини і те, що актуальність для соціології вивчення статистичних закономірностей аргументується у вигляді розгляду детермінованої і стохастичної (імовірнісної) складової у психології людини, аналізу механізму виконання емоційними формами психологічної діяльності людини ролі стохастичних регуляторів поведінки.

Проблема унікального та середнього.

Дослідник може прагнути знайти такі "обурення" у суспільному житті, такі її "переломні" точки системи, які свідчать або про її руйнування, або про зародження нової системи.

Природно, що з такій постановці завдання методи, створені задля пошук " середніх " закономірностей, які ховаються за спостережуваними фактами, тобто. статистичні методи, що перестають грати чільну роль.Пошук унікальних точок взагалі може не асоціюватись із пошуком закономірності.

Математична статистика як основа аналізу соціологічних даних: ознака, частота його значення, частотний розподіл, статистична закономірність у соціології.

Основними об'єктами вивчення математичної статистики є випадкові величини. Випадковими величинами в соціології є ознаки.Для кожної сукупності значень випадкової величини має бути визначена можливість, що, обстежуючи респондентів, соціолог зустріне значення з цієї сукупності.

При використанні математичної статистики можуть виникати такі складнощі:

1. Формування вибірки (інакше не можна буде екстраполювати на генеральну сукупність);

2. Не всі методи математичної статистики можна використовувати;

3. Відсутність суворих алгоритмів розв'язання багатьох практичних завдань.

Дескриптивна та індуктивна задачі аналізу даних.

Основні завдання розв'язувані математичною статистикою в соціології:

1. Дескриптивна (описова) – пошук статистичних закономірностей для вибірки, опис та пошук взаємозв'язків;2.

Індуктивна – узагальнення одержаних результатів не генеральної сукупності: статистична оцінка параметрів; перевірка статистичних гіпотез.

**Тема 2. Багатовимірне розгортання та індивідуальне багатовимірне шкалірування**

1. Постановка завдання; важливість врахування специфіки метрик окремих респондентів.
2. Спосіб обліку метрик в індивідуальному БШ.
3. Вид вхідних і вихідних даних, функції стресу в індивідуальному БШ.

Постановка завдання; важливість врахування специфіки метрик окремих респондентів. Спосіб обліку таких метрик в індивідуальному БШ. Вид вхідних і вихідних даних, функції стресу в індивідуальному БШ.

Багатовимірне шкалірування(БШ; англ. Multidimensionalscaling; (MDS)) – ряд пов’язаних між собою статистичних технік, що часто використовують в інформаційній візуалізації для дослідження схожості та відмінності у даних. БШ є особливим видом розміщення. БШ будується як матриця подібних елементів, після чого підписується розміщення кожного елементу у N-вимірному просторі, де через N позначають пріоритетність. Для достатньо малих N результат розміщень може бути представлений як графік чи візуалізований у 3D. БШ потрапляє в таксономію залежно від значення вхідних матриць.

Застосування включає наукову візуалізацію та глибокий аналіз даних в сфер когнітивних наук, інформаційних наук, психофізики, психометрики, маркетингу та екології. Нові застосування виникли з використанням незалежних безпровідних вузлів, які займають простір чи площу. БШ може застосовуватися як реальний підхід покращення використання часу для моніторингу та управління таким парком.

Більше того, БШ активно використовується у геостатистиці для моделювання просторової мінливості у графічних моделях, представляючи їх у вигляд точок у маловимірному просторі[2].

Одномірне розгортання. Обґрунтування необхідності переходу до простору довільної розмірності для успішного виконання завдання шкалірування. Модель ідеальної точки в багатовимірному випадку. Неметричне багатовимірне розгортання. Вид вихідних даних.

Функція стресу. Специфіка вихідних даних (наявність двох видів точок, що відповідають об'єктам і респондентам відповідно). Особливості інтерпретації результатів.

Суть багатовимірного шкалірування(БШ)

БШ - є одним із розділів прикладної статистики;

Термін «багатомірне шкалювання» відноситься до галузі аналізу даних. Йдеться про сукупність алгоритмів, що дозволяють реалізувати певний погляд на природу деяких завдань, що відповідає певній стратегії їх вирішення. Йдеться про завдання вивчення так званого простору сприйняття респондента.

Процес реалізації будь-якого алгоритму БШ – це процес вимірювання

БШ – інструмент наочного подання (візуалізації) вихідних даних.

Як оптимізований критерій якості методу в ньому використовується підсумована по всіх парах об'єктів відмінність вихідних (заданих) характеристик попарної близькості об'єктів від відповідних характеристик, обчислених у термінах шуканих координат об'єктів

Суть багатовимірного шкалірування (БШ)

БШ - «родина» геометричних моделей для багатовимірного представлення даних та відповідний набір методів для припасування таких моделей до реальних даних.

БШ - набір багатовимірних статистичних методів, призначених для визначення відповідності даних про близькість різними дистанційними просторовими моделями та для оцінки параметрів цих моделей.

Результати БШ

Результати багатовимірного шкалірування та аналізу інформації дозволяють отримати відповіді на запитання:

яка структура того чи іншого психологічного простору,

який ступінь подібності чи відмінності об'єктів,

які глибинні розрізняючі ознаки визначають суб'єктивну структуру зовнішньої дійсності, що сприймається.

В алгоритмі задається деякий критерій відмінності матриці близькостей та матриці відстаней. Цей критерій називається функцією стресу

ця функція змінюється від 0 до 1:

дорівнює нулю при повній схожості структур матриць і тим ближче до 1, чим це схожість менше.

Власне, термін «функція стресу» вперше був використаний Фарбалом для позначення введеної ним міри розбіжності між структурами материнки близькості та матриці відстаней у разі неметричного багатовимірного шкалювання.

Історія БШ - спроби подання стимулів у вигляді точок координатного простору відомі з давніх-давен.

Усвідомлення необхідності вирішення таких завдань відбулося приблизно у 50 60-х роках XX століття. Воно спричинило кардинальну зміну поглядів дослідників те що, якого роду дані можна отримувати від респондента, які їх вважатимуться адекватними, щодо якої інтерпретації даних можна досить впевненими у її коректності тощо.ВозникновениеБШможнорасцениватькакзакономерный результат протеканияпроцессовглобализации, информатизации, рынка и пр.

Історія БШ

І. Ньютон у книзі "Оптика" (1704 р.), описав приклад, в якому спектральні кольори представлені точками на колі, а відстані між ними відповідають розбіжності, що спостерігаються.

Дробіш (1846 р.) зумів розташувати звукові стимули на спіралі

Хеннінг (1916 р.) вдалося представити стимули, що відповідають запахам та смаковим відчуттям, на призмі та тетраедрі.

Література

Протягом 70-х - 80-х років у країні було опубліковано багато робіт, призначених для ознайомлення широких кіл соціологів з найбільш перспективними для вирішення соціологічних завдань математичними методами (див., наприклад, [Паніотто, Максименко, 1982], серію колективних монографій , випущених Інститутом соціології СРСР [ Інтерпретація та аналіз ..., 1987; Математичний аналіз та ..., 1989; Статистичні методи ..., 1979; Типологія та класифікація ..., 1982 ] , перекладену з англійської мови книгу [ Стенлі, 1976])

Сфера застосування БШ

БШ - корисно в антропології, педагогіці, географії, історії, психології, соціології, науках про поведінку, дослідження маркетингу.

БШ застосовується на дослідження соціальної структури організації, семантичної структури слів, логічної структури службових обов'язків.

Основний тип даних у БШ – міри близькості між двома об'єктами.

Міра близькості - це величина, визначена на парі об'єктів і вимірює, наскільки ці два об'єкти схожі. Часто зустрічаються такі заходи близькості, як коефіцієнти кореляції та спільні ймовірності.

Застосування БШ

Соціологу, що вивчає глядацькі уподобання, необхідно розуміти, що саме подобається людині в тій чи іншій передачі: симпатичний ведучий, гумор, зручний час виходу в ефір або щось ще.

Маркетолог повинен уявляти собі, чим керуються покупці, вибираючи сорт мила: наявністю в милі речовин, що пом'якшують шкіру, запахом або ціною і т.д.

Іншими словами, в соціології дуже важливою є завдання пошуку того простору, в якому люди уявляють собі цікаві дослідника об'єкти, визначаючи своє ставлення до них. Саме такий простір ми і називатимемо простором сприйняття.

Вираз знайти простір сприйняття означає вирішення кількох проблем: виявити кількість характеристик, якими керується респондент, визначаючи свою поведінку (або симпатії-антипатії), дати назви цим характеристикам, знайти координати кожного оцінюваного об'єкта у відповідному просторі

Традиційними способами збору даних при цьому є такі, які виходять, коли респондентів просять, до прикладу:

приписати кожному об'єкту число, що означає розташування об'єкта на осі (або відповідним чином проранжувати об'єкти), і вважають, що шкільною оцінкою об'єкта буде середнє значення приписаних йому чисел;

назвати кращий (найбільш підходящий посаду президента, подобається, часто купований тощо. буд.) об'єкт, і вважають, що шкільною оцінкою об'єкта буде частка респондентів, що вказали його, і т.д.

Традиційні методи виявлення простору сприйняття. Їх недоліки

Потенційні респонденти при визначенні свого ставлення до об'єктів, що розглядаються, керуються деякими властивостями цих об'єктів, дослідник перераховує в анкеті відповідні ознаки (наприклад, для сортів мила - ціна, запах, економічність і т. д.) і просить респондента відзначити в якою мірою кожна з цих властивостей присутня в тому чи іншому об'єкті.

Але при такому підході виникає одна з головних проблем, джерелом яких є «жорсткості» анкети.

Щоб уникнути зазначеного недоліку виявлення тих характеристик, у термінах яких респондент мислить собі якісь об'єкти, дослідник часто прямо звертається до респонденту з проханням вказати, які ж якості об'єктів його хвилюють

Традиційні методи виявлення простору сприйняття. Їх недоліки

Однак і такий підхід зазвичай призводить до некоректних результатів.

Відповідь може бути неадекватною через те, що респондент ніколи не замислювався над відповідним питанням і при відповіді сказав перше, що спало на думку, або ж (свідомо чи несвідомо) повторив якийсь штамп, згадав якийсь стереотип, механічно відтворив те, що багато разів чув у ЗМІ, і т.д.

приблизно до середини XX століття вважалося само собою зрозумілим, що зібрана соціологом статистична інформація має вигляд матриці (таблиці) «об'єкт-ознака».

Іншими словами, передбачалося, що в результаті збору даних кожен об'єкт, що вивчається, виявляється описаним сукупністю значень деяких ознак, тобто з формальної точки зору - представленим як точка деякого ознакового просторів

Традиційні методи виявлення простору сприйняття. Їх недоліки

Можливо, у відповідних методах збору даних є якийсь гносеологічний порок.

Ствердна відповідь на це питання (точніше, причин, що обумовлюють цю відповідь) і привела вчених до ідей БШ.

Відповідь на це питання якраз і полягала в пропозиції змінити метод збору даних (форму звернення до респондентів з питаннями) і потім, опрацювавши певним чином зібрану нетрадиційним способом інформацію, самому знайти латентний простір, що шукається, — і ті осі, які лежать у його основі, і відповідні координати об'єктів, що розглядаються.

Міра близькості

Основні типи завдань БШ

Завдання 1

Стиснення вихідного масиву даних з мінімальними втратами їх інформативності.

Якщо кількість аналізованих об'єктів велике (порядку кількох сотень, тисяч тощо. буд.), то вихідні дані видаються квадратною матрицею попарних близькість великої розмірності .

Розв'язання задачі БШ дозволяє перейти від форми вихідних даних типу "об'єкт - об'єкт" до більш поширеної та зручної для статистичної обробки формі вихідних даних типу "об'єкт - властивість", одночасно скоротивши обсяг масиву вихідних даних

Основні типи завдань БШ

Завдання 2

Верифікація геометричної конфігурації системи аналізованих об'єктів координатному просторі латентних змінних.

Йдеться ситуаціях, у яких з деяких змістовних (теоретичних) міркувань, які стосуються «фізичного» механізму досліджуваного явища, формулюються гіпотези про розмірності простору латентних змінних і тип геометричної конфігурації системи точок, що представляють аналізовані об'єкти у цьому просторі.

Результатом застосування БШ у завданнях цього є статистична перевірка (верифікація) згаданих гіпотез, їх уточнення.

## Тема 3. Проблеми формування вихідних даних і інтерпретації результатів у багатовимірному шкалірування

1. Роль соціолога при отриманні даних, вихідних для багатовимірного шкалірування, та інтерпретації його результатів.
2. Можливі способи одержання вихідних даних.
3. Безпосереднє отримання близькості від респондентів, класифікація відповідних способів опитування; проблеми, що постають при такому способі збору даних.
4. Приклади розрахунку матриці близькості на основі аналізу достатньо надійних даних іншого роду.

Роль соціолога при отриманні даних, вихідних для багатовимірного шкалірування, та інтерпретації його результатів. Можливі способи одержання вихідних даних. Безпосереднє отримання близькості від респондентів, класифікація відповідних способів опитування; проблеми, що постають при такому способі збору даних. Приклади розрахунку матриці близькості на основі аналізу достатньо надійних даних іншого роду.

Нижче наведені кроки для здійснення БШ дослідження:

Формулювання проблеми – як змінні ви хочете порівняти?

Пошук вхідних даних – респондентам задають ряд питань. Для кожної пари продуктів респондентів просять навести подібності (зазвичай за семизначною шкалою Лікерта, від дуже схожих до дуже різних). Ще один метод – «Метод даних за вподобаннями», якому респондентів просять надати перевагу якомусь товару, а не схожості між товарами.

Робота з БШ статистичними програмами - процедура БШ доступна в більшості статистичних програм. Існує вибір між метричним БШ (який дозволяє працювати з інтервалами чи даними про співвідношення рівня), і неметричним БШ (який працює з порядковими даними).

Відображення результатів та обґрунтування вимірів – статистична програма відобразить результати. Відображення буде здійснено по кожному продукту (зазвичай у двовимірному просторі). Наближення продуктів один до одного буде свідчити про те, наскільки вони схожі, або бажані, залежно від методу, що був застосований. Результати мають бути прокоментовані та інтерпретовані дослідником, що означає суб’єктивність у судженні та складність у роботі.

Використання формальних та неформальних методів при інтерпретації результатів багатовимірного шкалірування. Значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей.

Соціолог та БШ – простір сприйняття

Кожна зі згаданих характеристик відповідає одній координатній осі цього простору. Кількість таких характеристик число координатних осей називається його розмірністю

Слід зазначити ще один важливий момент: розмірність простору сприйняття має бути невеликою.

Можна лише відзначити, що розмірність 2, 3, 4 найчастіше розумно вважати невеликий. А розмірність 35, найімовірніше, не може бути визнана такою.

Основні типи завдань БШ

Завдання 3

Пошук та інтерпретація латентних змінних, що пояснюють задану структуру попарних відстаней. Цей тип завдань БШ передбачає як побудова допоміжних шкал, у системі яких потім розглядаються аналізовані об'єкти, а й змістовну інтерпретацію цих шкал як цілком певних характеристики.

Прикладні цілі БШ, по суті, не відрізняються від завдань, для вирішення яких залучається факторний, компонентний та латентно-структурний аналіз (різниця у формі завдання вихідної інформації).

Основні типи завдань БШ

Перша група – група координатних додатків. Організм людини певним чином реагує на стимули, дослідник повинен з'ясувати, які характеристики стимулів істотні для організму. Виявлені за допомогою БШ координатні осі вважаються суттєвими для організму характеристиками, а координати стимулів по цих осях інтерпретуються як значення цих характеристик.

У додатках типу «стиснення даних» користувач хоче стиснути складну матрицю близькості, уявити взаємини, що описуються цією матрицею, між стимулами в більш простому, зрозумілому вигляді. Якщо дані про подібність представлені у просторі, розмірність якого не вище трьох, то взаємини між стимулами можуть бути наочно зображені на одно-, дво- чи тривимірних

Основні типи завдань БШ

Останній тип додатків може бути названий "верифікація конфігурації". У таких додатках користувач починає з гіпотези про розмірність простору, яке має бути отримано, якщо застосувати до мір близькості неметричне багатовимірне шкалювання, і з припущення про вид знайденої при цьому конфігурації точок-стимулів.

Ця теорія передбачає, що при застосуванні неметричного БШ до кореляцій між шкалами в каталозі занять або в каталозі переваг професій шкали повинні лягти в двовимірному просторі вздовж шестикутника. В обговорюються питання застосування БШ для верифікації змін у завдання дослідження установок та розвитку, хоча там і не описані конкретні додатки.

Схема багатовимірного шкалювання. Етапи

Перший - це вибір безлічі об'єктів, що відповідає цілям дослідження.

Необхідно сконструювати матрицю попарних відмінностей (величин, зворотних близькостей) або матрицю суб'єктивних переваг, яка буде служити вхідною інформацією для наступного етапу.

Процедура опитування та вид оцінок повинні обиратися дослідником залежно від конкретної ситуації.

Схема багатовимірного шкалювання. Етапи

На другому етапі вирішується формальне завдання побудови координатного простору і розміщення в ньому точок-об'єктів таким чином, щоб відстані між ними, що визначаються за введеною метрикою, відповідали вихідним відмінностям.

Не потрібно ніяких відомостей про самі об'єкти, достатньо мати лише матрицю попарних відмінностей між ними.

Вибір методу вирішення цього завдання диктується цілями дослідження та типом вихідних даних.

Для побудови шуканого координатного простору використовується досить розроблений апарат лінійної чи нелінійної оптимізації.

Вводиться критерій якості ступінь розбіжності, що відображається між вихідними відмінностями (близьками) і результуючими відстанями (скалярними творами).

Схема багатовимірного шкалювання. Етапи

На другому етапі

Вигляд стресу залежить від умов, що накладаються на рішення конкретним автором. Знаходиться така конфігурація точок, яка давала б мінімальне значення цього стресу. Значення координат цих точок є рішенням задачі.

Використовуючи ці координати, ми будуємо геометричне уявлення об'єктів у просторі невисокої кількості вимірів.

Іноді ці подібності оцінюються в балах чи рангах. У деяких випадках задовольняються лише бінарною інформацією – відповідями "схожі — несхожі". Тоді матриця складається лише з нулів та одиниць. Часто вдаються до порівняння зі стандартним стимулом.

По черзі всі стимули беруться за стандартні, і випробуваного просять упорядкувати всі інші стимули за рівнем відмінності зі стандартним.

Кожне таке впорядкування є одним рядком, а всі разом вони утворюють повну матрицю, правда, в цьому випадку вона може бути несиметричною.

Схема багатовимірного шкалювання. Етапи

Третій етап

У методі тріад випробовуваному пред'являються одночасно три стимули, один з яких вибирається як стандартний, і запитується, який із двох, що залишилися, більше схожий на

стандартний. Підраховується, скільки разів було зазначено, що два стимули і більше схожі між собою, ніж інші стимули. Ця величина береться як міра подібності

Дистанційна модель для відмінностей

Аналогія між поняттям подібності у психології та поняттям відстані у геометрії.

Евклідова відстань має задовольняти наступним чотирьом аксіомам:

d (a, Ь) > 0,

d(a, a) = 0,

d(a, b) = d(b, a),

d (a, b)+d (b, c)≥ (a, c) - аксіома трикутника (нерівність трикутника)

Приклад БШ

Якщо об'єкти — автомобілі, ознаками можуть бути, наприклад, ціна, витрата бензину на милю, спортивність автомобіля.

Якщо об'єкти - місця роботи, то ознаками можуть бути престижність, заробітна плата, умови праці.

Нехай Xi та Xj — значення ознаки X об'єктів і та j відповідно. Наприклад, якщо об'єкти — автомобілі, а ознака — витрати бензину, то хі і хjозначатимуть витрати бензину цих автомобілів. Або якщо об'єкти – місця роботи, а ознака X – престижність, то хі та хj – престижність роботи і та j відповідно.

Дистанційна модель для відмінностей

Теоретичні величини можуть використовуватись у статистичній моделі для даних про відмінність. Ці теоретичні величини безпосередньо не спостерігаються і може бути оцінені за даними.

М. Річардсон [1938] запропонував почати з суб'єктивних суджень про відмінності об'єктів у парах і отримати ознаки, на яких ці судження засновані, а також значення стимулів за цими ознаками. Він запровадив завдання статистичної оцінки, звідки і з'явилося багатовимірне шкалювання, — завдання оцінки координат стимулів хік і хjк за мірами відмінностей.

Дистанційна модель відмінностей була прийнята на віру психологами-експериментаторами.

Дистанційна модель для відмінностей

Для ілюстрації перевірок аксіом розглянемо експеримент, описаний у [Rothkopf, 1957].

Е. Роткопф пред'являв випробуваному одні за одним два сигнали з абетки Морзе. Потім він просив випробуваного відповісти, чи ці сигнали одним і тим самим сигналом або двома різними. Усі можливі пари сигналів висувалися приблизно однакове число разів. Приблизно в половині пред'явлень стимулів і і j однієї парі першим пред'являвся стимул I, тоді як у другій

половині пред'явлень — стимул j. Число пред'явлень пари стимулів (i, j), у яких випробуваний дав відповідь «різні», може розглядатися як міра слухової різниці між стимулами і і j.

Відповідь «різні» була дана при пред'явленні двох сигналів азбуки Морзе Е в 3% випадків, а при пред'явленні двох сигналів Р — 17%. Психологи проінтерпретували такі результати як порушення аксіоми відстані

Дистанційна модель для відмінностей

Взяті разом, аксіоми (1.1) і (1.2) ведуть до припущення, що у жодному разі за умови пред'явлення пари різних сигналів відповідь «різні» ні з'являтися рідше, ніж за пред'явленні пари ідентичних сигналів.

У даних Роткопфа є невелике порушення цього принципу. При пред'явленні двох сигналів Р відповідь «різні» було дано у 17 % випадків. Коли сигнал Х слідував за сигналом, відповідь «різні» було дано лише у 16 % випадків. Психологи проінтерпретували ці результати як свідчення того, що дані про різницю не задовольняють першим двом аксіомам відстані.

Якщо нерівність трикутника не виконано, але виконані три інші аксіоми, дані можна перетворити таким чином, що нерівність трикутника буде виконуватися.

Висновки, отримані з експериментальних досліджень оцінок відмінності, неможливо знайти розширені інші типи заходів близькості. Проте перевірки аксіом, проведені психологами, можуть бути застосовані до будь-якого можливого індексу близькості. Наприклад,

залишається мірою близькості, придатною для застосування у багатовимірному шкалюванні. Аналогічно аксіоми (1.1) - (1.4) дають основи оцінки придатності будь-якої міри близькості для використання її в БШ.

**Тема 4. Канонічний аналіз. Загальне уявлення про методи, які засновані на моделях частот**

1. Загальне уявлення про моделювання частот таблиці спряженості.
2. Мультиплікативні та адитивні моделі частот. Роль логарифмування мультиплікативної моделі.
3. Поняття зв'язку між двома групами ознак.
4. Основне завдання канонічного аналізу. Послідовність канонічних коефіцієнтів кореляції. Принципи їх отримання на основі аналізу таблиці спряженості.

Загальне уявлення про моделювання частот таблиці спряженості. Змістовне розуміння таких моделей, їх роль для соціолога. Мультиплікативні та адитивні моделі частот. Роль

логарифмування мультиплікативної моделі. Можливість різного розуміння як сенсу розглянутих вкладів, так і того "середнього" рівня, з яким порівнюються спостерігаються частоти в процесі їх моделювання.

Канонічний кореляційний аналіз – один із мето- дів багатовимірного аналізу даних. Це найбільш узагальнена форма аналізу кореляцій, яка дозволяє досліджувати взаємозв'язок між двома множинами змінних, на відміну від факторного аналізу, який застосовують для встановлення зв'язків усередині однієї множини змінних. Метод канонічного аналізу відносно молодий. Уперше його ідею було опубліковано американським економістом Гарольдом Хотеллінгом (H.Hotelling) у журналі Біометрика у 1936 р. Однак активно теорія канонічного аналізу розроблялася вже у 70-ті рр. XX ст. з розвитком відповідного програмного забезпечення. На сьогоднішній день канонічний аналіз використовується у марке- тингових, економічних, природничих, медичних дослідженнях Поняття зв'язку між двома групами ознак. Послідовність канонічних коефіцієнтів кореляції. Принципи їх отримання на основі аналізу таблиці спряженості.

Використання канонічної кореляції в аналізі таблиць спряженості. Моделі частот, що відповідають канонічному аналізу. Зв'язок канонічних коефіцієнтів кореляції з критерієм «хі-квадрат". Загальне уявлення про оцифрування значень номінальних ознак. Необхідність сполучення моделі, закладеної в конкретному методі оцифровки, з вмістом розглянутої задачі. Приклад моделі такого роду – модель, використовувана в методі шкалірування, званому методом послідовних розбивок. Канонічний аналіз як метод оцифровки і метод вимірювання зв'язку між двома номінальними ознаками зі "спільними альтернативами". Моделі частот, що відповідають канонічному аналізу. Побудова соціологічних індексів за допомогою техніки канонічного аналізу. Вирішення проблеми зважування складових індекс ознак.

З даних, що містяться в матриці близькості, отримують оцінки координат стимулів хік та хjк.

Але як отримати міру близькості пари стимулів і як спланувати дослідження для збору таких заходів?

В основному існують чотири класи заходів близькості:

прямі оцінки відмінності,

умовні ймовірності,

спільні ймовірності

індекси відмінності профілів

ПРЯМІ ОЦІНКИ ВІДМІННОСТІ

Необхідно побудувати вибірки стимулів піддослідних, здійснити вибір завдання оцінки і планування інструментарію (вопросника) подання цього завдання піддослідним.

Першим етапом отримання прямих оцінок відмінностей є наскільки можна найточніше визначення генеральної сукупності людей, які мають збирати оцінки, і генеральної сукупності стимулів, які необхідно обирати.

ПРЯМІ ОЦІНКИ ВІДМІННОСТІ

Щойно генеральна сукупність стимулів визначено, дослідник має спробувати побудувати випадкову чи стратифіковану випадкову вибірку стимулів.

Якщо деякі зі стимулів генеральної сукупності невідомі багатьом випробуваним, може виявитися необхідним звузити обрану генеральну сукупність стимулів до відомої всім випробуваним частини.

Піддослідні можуть обґрунтовано оцінити ті стимули, з якими добре знайомі.

ПРЯМІ ОЦІНКИ ВІДМІННОСТІ

При відборі стимулів для стратифікації вибірки та забезпечення репрезентативності побудованої вибірки з генеральної сукупності стимулів дуже корисні класифікації цих стимулів.

Наприклад, для дослідження сприйняття професій можна використовувати класифікаційну систему, розроблену в Словнику назві професій служби зайнятості США

Якщо стимули — цілі навчання у конкретній предметної області, то стратифікацію можна з урахуванням тим, відповідних назв глав у підручнику з цього предмета.

Як правило, вибірка повинна містити щонайменше п'ять стимулів на кожну очікувану координатну вісь.

Важливість відбору стимулів

Процес відбору стимулів може істотно вплинути на отримане рішення

Якщо у вибірку включені лише «білі комірці», то не слід очікувати появи координатної осі, що відповідає безпеці професії, оскільки всі професії «білих комірців» безпечні.

Якщо вибірка містить професії «синіх комірців», пов'язані з великим ризиком, такі, як шахтар або металіст, можна очікувати появи осі безпеки занять, оскільки професії у вибірці різняться з їхньої безпеки.

Важливі етапи

Важливими етапами є визначення генеральної сукупності суб'єктів (випробуваних) та побудова вибірки з неї.

Якщо деяка частина генеральної сукупності суб'єктів не знайома зі стимулами, які треба досліджувати, то дослідник може постати перед необхідністю обмежити генеральну сукупність тими, хто може обґрунтовано оцінити всі стимули.

Стратифікація, генеральна сукупність за такими змінними, як стать, вік та рівень освіти, може бути гарантією репрезентативності вибірки.

Важливі етапи

Застосовується таке емпіричне правило: якщо відмінність пари об'єктів оцінюється шляхом усереднення оцінок різних піддослідних, число М усереднених оцінок кожної пари стимулів повинно

Завдання для оцінки

Після побудови вибірок стимулів та випробуваних експериментатор повинен вирішити, який тип завдання для оцінки відмінностей слід дати випробуваним.

У літературі зустрічаються різні види завдань. Тут описано чотири з них:

оцінка величини відмінності,

категоріальна оцінка,

графічна оцінка

категоріальне сортування.

Завдання для оцінки

величина відмінності

У завданні з оцінки величини відмінності одна пара стимулів вибирається як стандарт.

Випробуваний повинен приписати кожній парі, що оцінюється, число, що показує ступінь відмінності цієї пари щодо відмінності стандартної пари.

Наприклад, якщо випробуваний вважає, що оцінювана пара вдвічі більш різна, ніж стандартна пара, він припише оцінюваної парі число 2.

Якщо ж випробуваний вважає, що відмінність оцінюваної пари становить дві третини від відмінності стандартної пари, він припише оцінюваної парі число 2/3.

Завдання для оцінки

величина відмінності

категоріальні оцінки

Для отримання прямих оцінок відмінностей найчастіше вдаються до категоріальних оцінок. У цьому випадку випробуваному пред'являють пару стимулів таким чином:

Йому дається завдання вказати, наскільки схожими (або різними) він вважає пару стимулів і відзначити відповідну категорію на оцінці. Часто добрі результати дає шкала, що містить від шести до дев'яти категорій.

Завдання для оцінки  категоріальні оцінки

Зазвичай кожній категорії приписується ціле число, і відповіддю досліджуваного вважається число, приписане зазначеній категорії.

У наведеному прикладі категоріям можуть бути приписані цілі числа, наведені в дужках нижче шкали оцінок. Показаний у цьому прикладі відповідь отримає оцінку 4. В

якості оцінки відмінності пари береться середня арифметична оцінка, приписана їй усіма піддослідними.

Метод шкали графічної оцінки

Метод шкали графічної оцінки дуже схожий метод категоріальної оцінки. І тут випробуваному пред'являється пара стимулів так:

Випробуваний повинен, як показано вище, перекреслити шкалу у точці, відстань від якої до лівого краю шкали відповідає різницю між двома оцінюваними стимулами.

У цьому прикладі ця відстань дорівнює 6 см. Якщо необхідно об'єднати відповіді кількох суб'єктів, то як оцінка міри відмінності двох стимулів береться середнє арифметичне (рідше медіана). Вимірювати на графіку відповіді кожного випробуваного лінійкою дуже незручно.

Метод шкали графічної оцінки

Другий варіант методу отримання категоріальної оцінки - категоріальне сортування.

У цій процедурі кожна пара стимулів представлена окремою карткою.

Випробовуваний поміщає кожну пару до однієї з кількох упорядкованих категорій. Він має віднести пари дуже схожих стимулів до найнижчої категорії. Найвища категорія містить пари, які зовсім не схожі.

Для того, щоб в одній категорії не виявилося занадто великої кількості пар, експериментатор може встановити, скільки категорій слід використовувати випробуваному (часто від шести до дев'яти) і яку частину пар стимулів поміщати в кожну з них.

Метод шкали графічної оцінки

Оцінки відмінності за категоріальним сортуванням виходять так само, як і в задачах за категоріальною оцінкою. Кожній категорії приписується ціле число, а оцінка суб'єкта — число, яке відповідає категорії, до якої вміщено пару. Для кожної пари відмінністю вважається середнє арифметичне приписаних їй оцінок.

Для отримання прямих оцінок відмінностей дослідник може вибрати, принаймні, чотири типи завдання для оцінки: оцінка величини, категоріальна оцінка, графічна оцінка та категоріальне сортування.

Вибрати один із цих методів можна, розглянувши практичні проблеми, що виникають при використанні кожного з них, і оцінивши здібності та бажання досліджуваних виконувати різні завдання. Для оцінки цих факторів часто потрібне невелике пілотажне (пробне) дослідження.

Планування інструментарію

Визначивши тип завдання, експериментатор повинен продумати, як уявити його випробуваним.

Пари стимулів, які потрібно оцінити, слід упорядкувати в логічну послідовність.

Досліджувані повинні отримати відповідну інструкцію. Потім експериментатор повинен вирішити, чи буде використовувати повний план, коли кожен піддослідний оцінює всі можливі пари стимулів, або неповний, коли кожен піддослідний оцінює тільки одне підмножина пар стимулів.

Планування інструментарію

проблеми: упорядкування пар стимулів, підготовка інструкцій та використання неповних планів.

**Тема 5. Логлінейний аналіз**

1. Причини відхилення спостережуваних частот від їхніх середніх значень, тобто відмінності реального розподілу від рівномірного.
2. Неможливість отримання нового знання на основі аналізу рівномірного розподілу (суть аналізу даних – вивчення змін, порівняння показників різного роду).
3. Моделі частот, що відповідають логлінейному аналізу.
4. Насичена модель.
5. Мета переходу до логарифмів частот. Сенс вкладів різної розмірності.

Логлінейний аналіз - метод багатовимірного статистичного аналізу для вивчення таблиць спряженості. Логлінейний аналіз дозволяє статистично перевіряти гіпотезу про систему одночасно мають місце парних і множинних взаємозв'язків в групі ознак, виміряних за номінальними шкалами. Багатовимірний статистичний аналіз.

Причини відхилення спостережуваних частот від їхніх середніх значень, тобто відмінності реального розподілу від рівномірного. Неможливість отримання нового знання на основі аналізу рівномірного розподілу (суть аналізу даних – вивчення змін, порівняння показників різного роду).

Моделі частот, що відповідають логлінейному аналізу. Насичена модель. Мета переходу до логарифмів частот. Сенс вкладів різної розмірності.

Гіпотези про взаємозв'язок ознак. Їх роль при побудові моделей частот. Проблема формування таких гіпотез. Роль критерію "хі-квадрат" при використанні логлінейного аналізу.

Розрахунок коефіцієнтів логлінейной моделі для двовимірного випадку. Відносини переважання. Інтерпретація коефіцієнтів через відносини переважання (для моделі довільної розмірності).

Порівняння логлінейного аналізу з номінальним регресійним і дисперсійним аналізом, а також з методом послідовних розбивок. Порівняння здійснюється на змістовному рівні.

Різне розуміння залежної ознаки: кількісна ознака в дисперсійному аналізі, кількісна або номінальна – в номінальному регресійному і частота, що стоїть в клітці багатовимірної таблиці спряженості, – в логлінейном аналізі. Різні можливості пошуку поєднань значень предикторів: перевірка гіпотез про наявність багатовимірних зв'язків у логлінейном аналізі і можливість пошуку найбільш дієвих поєднань в методі послідовних розбивок і регресійному аналізі, заздалегідь заданий набір поєднань значень предикторів в дисперсійному аналізі.

Упорядкування пар стимулів

існують два чинники, що впливають оцінки піддослідних, та його слід враховувати під час виборів логічної послідовності пар стимулів.

Порядок пред'явлення стимулів парі (т. е. Гарвард — Йейл чи Йейл — Гарвард) явно впливає оцінки подібності цих двох стимулів. Такий вплив називається просторовим ефектом.

Просторові ефекти для даного стимулу збалансовані, якщо в одній половині включають цей стимул пар він є першим, а в іншій - другим. Пари стимулів повинні пред'являтися таким чином, щоб для кожного стимулу, що оцінюється, просторові ефекти були збалансовані.

упорядкування пар стимулів

Тимчасові ефекти - це ефекти, пов'язані з упорядкуванням пар стимулів у списку оцінюваних пар.

Тимчасові ефекти для даного стимулу збалансовані, якщо пари, які включений цей стимул, розташовані в списку рівномірно. В ідеалі, тимчасові ефекти повинні бути збалансовані для кожного стимулу.

Р. Росс описує метод визначення такого розташування та впорядкування пар стимулів, коли збалансовані як просторові, і тимчасові ефекти. Це називається упорядкуванням Росса.

Випадкове впорядкування

Альтернативою впорядкування Росса є випадкове впорядкування, коли пари стимулів висуваються у випадковому порядку.

Для кожної пари експериментатор за допомогою якогось випадкового процесу типу кидання монети вирішує, який із двох стимулів з'явиться першим.

У цій процедурі рішення про упорядкування стимулів складається з двох етапів упорядкування пар та випадкового вибору кожного члена пари. Випадкове впорядкування пар не гарантує балансування просторових та часових ефектів.

Третій метод — метод стандарту, що чергується, — дуже простий для реалізації.

Якщо є I стимулів, пронумерованих 1, …, і, …, I, то завдання з упорядкування пар ділиться на (I - 1) розділів.

У і-му розділі стандартом є стимул I.

Завдання випробуваного - дати в і-му розділі оцінку подібності між стандартом і кожним стимулом, пронумерованих від (i+1) до I.

Метод стандарту, що чергується

На рис. показані дві сторінки з запитальника, складеного за методом стандарту, що чергується.

Метод стандарту, що чергується

Зазвичай номерів розділів у запитальнику немає. На малюнку вони включені лише з метою обговорення.

Стимули пронумеровані так, як показано у верхній частині малюнка.

У розділі I стимул 1 (Гарвард) порівнюється з кожним із шести стимулів, що залишилися.

У розділі II стимул 2 (Йейл) порівнюється з усіма стимулами крім першого. Порівнювати Йейл з першим стимулом немає необхідності, оскільки це порівняння було проведено у розділі I.

У розділі III стимул 3 (Аітіок-коледж) порівнюється зі стимулами 4-7. Немає необхідності порівнювати цей стимул з жодним з перших двох.

У цьому вся методі не збалансовані ні просторові, ні тимчасові ефекти.

Оскільки впорядкування Росса балансує просторові та тимчасові ефекти, то це найкращий план. Наступним за якістю є випадкове впорядкування.

Якщо експериментатор не знає, що просторові та тимчасові ефекти будуть мінімальними, то основою для рекомендації методу стандарту, що чергується, буде тільки його раціональність.

Інструкції

Вчені припускають, що оцінки випробуваного можуть змінитися в залежності від того, що він очікує побачити у наборі стимулів.

Інструкції можна використовувати для стандартизації очікувань кожного випробуваного, якщо попросити випробуваних прочитати весь список стимулів після читання інструкції, але до початку оцінювання.

Якщо читати весь список стимулів занадто незручно, то піддослідні можуть прочитати представницьку вибірку з нього, щоб зрозуміти, які стимули вони оцінюватимуть.

Інструкції

Інструкції можуть обмежувати характеристики, які експериментатор вважає за важливі для оцінок.

Піддослідних можна попросити не брати до уваги характеристики, що не належать до справи,

наприклад, такі, як довжина слова, що означає стимул.

Деякі з характеристик, за якими змінюються стимули, іноді не мають відношення до цілей дослідника.

Випробуваних можна попросити оцінювати навчальні заклади у припущенні, що вони будуть прийняті в кожний з них та отримають повну фінансову підтримку.

Інструкції

Таке припущення дозволяє випробуваному ігнорувати дві можливі характеристики: труднощі вступу та оплата освіти.

Дослідник може вказати деякі з характеристик, які слід розглянути.

Наприклад, випробуваного, що оцінює навчальні заклади, можна попросити розглянути лише ті характеристики, які вплинули на рішення навчатися в одному з них, а не в іншому.

Неповні плани

Неповні плани

Як же розділити пари на підмножини?

У роботах (MacCallum, 1979) висловлено припущення, що випадкове поділ пар на підмножини не гірше за будь-який інший спосіб.

Кожна пара повинна з'явитися в однаковій кількості підмножин. Якщо можливо, кожен із стимулів повинен з'явитися принаймні в одній парі кожної підмножини.

Підмножини, що частково перетинаються, дають хороші результати, але вони не є необхідними. Оцінки координат стимулів повинні бути адекватними, доки кожну пару оцінює не менше ніж M=40K*/(I-1) випробуваних

Висновки

Для використання оцінки відмінностей у дослідженні із застосуванням МШ повинні при плануванні дослідження розглянути такі питання.

По-перше, слід визначити сукупність стимулів і сукупність піддослідних та створити план для побудови адекватних вибірок тих та інших.

По-друге, потрібно вибрати завдання для оцінки. Хороші результати дають методи оцінки величини, категоріальної оцінки, графічної оцінки та категоріального сортування.

По-третє, досліднику необхідно продумати інструментарій пред'явлення завдання випробуваним.

При розробці такого інструментарію експериментатор упорядковує пари стимулів таким чином, щоб просторові та часові ефекти були збалансовані.

З інструкції всі випробувані мають отримати однакове уявлення про набір стимулів, які оцінюватимуться.

Якщо необхідно, в інструкції має бути сказано, що випробуваним слід ігнорувати характеристики стимулів, що не належать до справи.

Якщо число стимулів велике, експериментатору слід розробити неповний план, у якому кожен випробуваний оцінюватиме лише невелике підмножина пар стимулів.

Умовні та спільні ймовірності

Як міри близькості стимулів використовувалися різні типи матриць умовної та спільної ймовірності.

Умовна ймовірність - ймовірність однієї події за умови, що інша подія вже сталася.

- число результатів, що сприяють спільній події подій E і E1, - число результатів, що сприяють появі події E1. Знаючи числа елементарних результатів можна розрахувати умовну ймовірність.

Умовні та спільні ймовірності

Імовірність спільної появи двох незалежних подій E1 та E2 дорівнює добутку їх ймовірностей.

n(E1) – число наслідків сприятливих події E1; n(E2) – число наслідків сприятливих події E2; n1 – число наслідків сприятливих та несприятливих події E1; n2 - кількість наслідків сприятливих і несприятливих події E2.

## Тема 6. Причинний аналіз. Стратегія аналізу структури взаємозв'язків ознак

1. Роль статистичних методів при вивченні причинних відносин.
2. Граф причинних зв'язків. Структурні коефіцієнти. Вхідні (зовнішні, незалежні) і вихідні (внутрішні, залежні) змінні.
3. Правила редукції причинних схем та формування рівнянь.
4. Повторення принципів побудови часткових коефіцієнтів кореляції і регресії.

Поняття причини в соціології. Принципова неможливість повністю його формалізувати. Роль статистичних методів при вивченні причинних відносин.

Граф причинних зв'язків. Структурні коефіцієнти. Вхідні (зовнішні, незалежні) і вихідні (внутрішні, залежні) змінні. Правила редукції причинних схем та формування рівнянь.

Повторення принципів побудови часткових коефіцієнтів кореляції і регресії. Важливість для соціолога вивчення відповідних зв'язків. Різниця між статистичним та причинним зв'язком. Поняття "помилкової" кореляції. Основні причинні схеми, що призводять до їх появи.

Координуючий шлях. Його ефект. Обчислення коваріацій (кореляцій) між будь-якими двома ознаками на основі графа зв'язків. Вивчення статистичних зв'язків на основі причинних схем як основне завдання причинного аналізу.

Структурні рівняння. Обчислення структурних коефіцієнтів. Їх зв'язок з частковими коефіцієнтами регресії. Основна теорема причинного аналізу. Її роль у вивченні статистичних залежностей.

Поняття допоміжної теорії вимірювань Блейлока. Причинний аналіз як концептуальний підхід до вивчення соціальних явищ.

Проблема формалізації завдання вивчення причинно-наслідкових відносин в соціології. Поняття структури багатовимірної випадкової величини. Формування узагальнених показників на базі аналізу структури зв'язків ознак. Комплексне використання декількох методів вивчення зв'язків між ознаками для вирішення соціологічних задач (аналіз структури випадкової величини; факторний і дисперсійний аналіз; пошук детермінуючих поєднань значень предикторів).

Умовні та спільні ймовірності

У дослідженнях, де застосовуються умовні ймовірності, як міра подібності стимулів (чи подій) служить ймовірність того, що стимул (подія) j зустрічається за наявності стимулу (події) I.

У дослідженнях, де застосовуються спільні ймовірності, як міра подібності стимулів (подій) і та j береться ймовірність того, що стимули (події) і та j зустрічаються разом. Такі ймовірності розглядаються як заходи подібності, а чи не відмінності.

Матриці переходу несиметричні. І в даному випадку найпростіший спосіб побудови симетричної матриці близькості – створити нову матрицю з елементами

$\delta ij = \delta ji = rij + pji$.

На рис. показана гіпотетична матриця переходу та отримана з неї матриця подібності.

Умовні та спільні ймовірності

Рядки відповідають спеціальностям у коледжі під час вступу студентів, а стовпці — спеціальностям тих самих студентів після закінчення.

Кожен елемент pij - частка студентів, які оголосили під час вступу спеціальність I, а закінчили коледж зі спеціальності J.

Наприклад, елемент у рядку 2, стовпець 4, показує, що 40% студентів, які оголосили під час вступу своєю спеціальністю біологію, закінчували як педагоги.

Діагональні елементи P і A під час аналізу будуть ігноруватися.

A - це матриця подібності, складена з величі $\delta ij = \delta ji = pij + pji$.

Умовні та спільні ймовірності

Дослідження соціальної взаємодії можуть породити умовні матриці ймовірностей.

Рядки та стовпці відповідають людям чи громадським організаціям.

Елемент pij у матриці умовної взаємодії - частка взаємодій з ініціативи особи I (або організації I) з особою (або організацією) J.

Умовні та спільні ймовірності

Зазвичай такі матриці не симетричні.

Найпростішим способом побудови симетричної матриці близькості A буде утворення елементів δij = δji = rij + pji.

При цьому передбачається, що стимул I більш схильний розпочинати взаємодії зі стимулами, схожими на нього. Тому pij та δij відображають подібність між взаємодіючими особами (організаціями) I та J.

Умовні та спільні ймовірності

При побудові матриць подібності, що є вихідними даними для МШ, можуть бути використані інші типи матриць умовної ймовірності, якщо ці умовні ймовірності побудовані так, що вони відображають близькість між об'єктами.

Оскільки матриці умовної ймовірності практично ніколи не бувають симетричними, дослідник повинен винайти спосіб побудови симетричних матриць близькості асиметричних матриць умовної ймовірності.

Умовні та спільні ймовірності

Однак, при побудові симетричних матриць з несиметричних матриць умовної ймовірності може бути втрачена важлива інформація.

Асиметрії можуть виникнути через те, що відносини між двома об'єктами суттєво асиметричні. При побудові симетричних матриць дослідник може втратити інформацію про невзаємному характері взаємовідносин об'єктів, що міститься в ймовірностях і тому отримане МШ-рішення не буде відображати цю асиметрію.

Умовні та спільні ймовірності

Наприклад, матриця P на рис. показує, що можливість переходу від медицини до бізнесу 0,40, а від бізнесу до медицини набагато менше - 0,05. Ця асиметрія не відображається і не буде відображена у відстані між медициною та бізнесом на багатовимірній шкалі чотирьох спеціальностей.

Якщо симетрична матриця близькості породжена з умовних ймовірностей, то багатовимірна шкала, що результує, може відображати багато важливих властивостей стимулів, але не ті, які пов'язані з асиметричністю.

Спільні ймовірності

Спільні ймовірності різного типу можуть бути індикаторами подібності між парами об'єктів. Так як за визначенням спільні ймовірності симетричні, вони завжди задовольняють принаймні одній вимогі до матриці для багатовимірного шкалювання.

Спільні ймовірності

Дослідження соціальної взаємодії, результатом яких можуть бути несиметричні матриці умовних ймовірностей, іноді дозволяють також отримати симетричні матриці спільних ймовірностей.

Рядки та стовпці в матриці спільної взаємодії представляють взаємодіючі елементи.

Елемент рij у матриці спільної взаємодії - частка взаємодій, що включають елементи I та J.

Спільні ймовірності

Так як матриця симетрична, вона може служити як матриця подібності для аналізу із застосуванням багатовимірного шкалювання. Передбачається, що й елементи I і J дуже схожі, всі вони більш схильні взаємодіяти.

Такі матриці спільної взаємодії є лише одним можливим типом матриць спільної ймовірності для МШ-аналізу.

Дослідження, проведене Американською психологічною асоціацією

Було підраховано кількість членів АПА, які перебувають одночасно у кожній із усіх можливих пар її підрозділів.

Дослідження призвело до великої матриці спільних ймовірностей Р, що містить один елемент для кожної пари підрозділів.

Елемент рij представляє частку членів, які перебувають одночасно у підрозділах і та j. Матриця спільних ймовірностей була піддана МШ-аналізу для одержання просторового уявлення структури підрозділів АПА.

Така матриця називається матрицею спільної встречаемості, оскільки кожен її елемент відповідає спільної встречаемості двох подій - членству однієї й тієї ж особи у двох підрозділах.

Дослідження, проведене Американською психологічною асоціацією

Матриця спільних ймовірностей у дослідженні АПА було побудовано за записами асоціації. Матриці спільної зустрічальності можуть бути отримані іншим способом.

Випробуваним було дано перелік характеристик характеру. Досліджувані присвоювали кожну з рис тому людині зі списку своїх знайомих, якого ця риса найкраще характеризувала.

Міра подібності рис і і j виходила шляхом підрахунку загальної кількості (або частки) знайомих, яким приписані обидві риси.

Наприклад, якщо піддослідний назвав 100 знайомих і 47 з них охарактеризував як проникливих та інтелігентних, то мірою подібності між рисами «проникливість» і «інтелігентність» може бути число 47 або відповідний дріб 0,47.

Метод Розенберга

Метод Розенберга пов'язаний з наступним випробуваним завданням:

розсортувати риси характеру на категорії, причому кожна категорія відповідає одному із знайомих випробуваного.

Категоріальна сортування може бути застосована до стимулів будь-якої природи, тобто випробуваним можна дати завдання розсортувати набір стимулів на номінальні категорії таким чином, щоб стимули в будь-якій з категорій були в певному сенсі схожі один на одного і відрізнялися в тому ж сенсі від стимулів інших категоріях.

Завдання Розенберга

Завдання Розенберга, яке тут називається завданням сортування стимулів, відрізняється від завдання категоріального сортування, описаного в параграфі прямих оцінок відмінностей.

По-перше, при категоріальному сортуванні піддослідний повинен розділяти за категоріями пари стимулів, а чи не поодинокі стимули.

По-друге, при категоріальному сортуванні категорії впорядковані за схожістю від категорії найбільш схожих стимулів до категорії найменш схожих. У разі сортування стимулів категорії повністю номінальні.

Висновок

Матриці ймовірностей різного виду є альтернативою прямих оцінок відмінностей МШ-дослідженнях.

Однією з таких альтернатив є матриці умовних ймовірностей, у тому числі матриці переходу.

Матриці умовних ймовірностей несиметричні, необхідно до застосування до них МШ побудувати із матриці умовних ймовірностей симетричну матрицю близькості.

Може бути втрачено частину інформації про асиметричність відносин між об'єктами.

Висновок

Інша альтернатива прямим оцінкам відмінностей - матриці спільних ймовірностей.

Один із методів отримання матриць спільних ймовірностей — сортування стимулів, запропоноване С. Розенбергом.

Метод сортування стимулів може застосовуватися для шкалювання великої кількості стимулів без навантаження випробуваних.

МІРІ ВІДМІННОСТІ ПРОФІЛІВ

Профіль – це просто набір кількісних ознак об'єкта.

Якщо під об'єктом розуміється людина, то кількісними ознаками може бути його оцінки з різних тестів.

Якщо об'єкти — міста, то профіль кожного міг би складатися із записів температури в 10 різних моментів року.

У табл. наведено відсоток безробітних у п'яти сферах зайнятості за п'ять різних років. Кожен рядок таблиці – профіль оцінок відповідної сфери зайнятості.

У цьому параграфі припускатимемо, що профілі оцінок об'єктів розташовані у вигляді матриці, такої, як табл., де рядки відповідають об'єктам, а стовпці — ознакам. Якщо матриця V містить профілі, то vik - к-й знак об'єкта I.

Було запропоновано різні заходи подібності профілів, але найпоширеніша міра — відстань

У деяких випадках відмінність визначається як квадрат цієї величини:

У цих виразах vik і vjk є значення k-го ознаки в профілях і-го і j-го об'єктів. Мірою подібності профілів об'єктів і і j є j (або j).

(3.1) сума виходить шляхом складання квадратів різниць між відповідними елементами в і-й і j-й рядках матриці V.

Наприклад, міра відмінності у перших двох об'єктів (працюючі за наймом, робітники в гірничодобувній промисловості) у табл. 3.1 буде

Відмінності будуть підраховуватися для елементів стандартизованої по стовпцях матриці Z, а не V.

Середні та дисперсії шкал суб'єктивних оцінок та багатьох психологічних тестів часто розглядаються як довільні. Якщо профілі складаються з таких змінних, доречна стандартизація даних по стовпцях.

У психологічних дослідженнях дані часто бувають центрованими або стандартизованими рядками.

Матриця даних називається центрованою за рядками, якщо середнє елементів кожного рядка дорівнює 0,0. Вона називається стандартизованою за рядками, якщо середнє елементів кожного рядка дорівнює 0,0, а дисперсія елементів кожного рядка дорівнює 1,0.

Немає необхідності окремо розглядати ці перетворення даних. Матриця відмінностей профілів A, побудована на матриці центрованих по стовпцях даних, дорівнюватиме матриці відмінностей профілів, визначеної за початковими величинами.

Аналогічно матриця відмінностей профілів V, визначена на матриці даних з подвійним центруванням, дорівнюватиме матриці відмінностей профілів, визначеної на матриці, центрованої по рядках.

Розбіжність профілів, визначена в (3.1), може використовуватися фактично для будь-яких числових профільних даних.

Однак для отримання профілю оцінок кожного стимулу досліднику необхідно знати суттєві ознаки стимулів. Більш того, дослідник повинен мати можливість виміряти кожну з суттєвих ознак, щоб ці вимірювання включити в профіль. З іншого боку, щоб використовувати прямі оцінки відмінностей або ймовірні міри близькості, дослідник не повинен вміти визначати важливі ознаки або міряти кожен з них.

Тому в тих випадках, коли важливі ознаки стимулів погано зрозумілі або важко отримати незалежні виміри за кожною з ознак, прямі оцінки відмінностей і ймовірнісні заходи краще заходів відмінності профілю

**Тема 7. Завдання розпізнавання образів. Поняття автоматичної класифікації об'єктів**

1. Класифікація як один із фундаментальних процесів у науці. Ознаковий простір.
2. Задача класифікації як пошук згущення точок – моделей об'єктів в ознаковому просторі.
3. Загальне уявлення про завдання розпізнавання образів (синоніми: образ, клас, кластер, таксон; неоднозначність трактування термінів в літературі).
4. Виділення завдань: пошук класів, опис класів, визначення найбільш ефективної системи ознак. Роль наявності або відсутності навчальної вибірки.

Класифікація як один із фундаментальних процесів у науці. Ознаковий простір. Задача класифікації як пошук згущення точок – моделей об'єктів в ознаковому просторі.

Клас - об'єктом мають загальні властивості. Для об'єктів одного класу передбачається наявність «схожості». Для завдання розпізнавання може бути визначено довільну кількість класів, більше.

Класифікація - процес призначення міток класу об'єктів, відповідно до деякого опису властивостей цих об'єктів. Класифікатор - пристрій, який в якості вхідних даних отримує набір ознак об'єкта, а в якості результату видає мітку класу.

Верифікація - процес зіставлення примірника об'єкта з однією моделлю об'єкта або описом класу.

Іншими словами, розпізнавання образів можна визначити, як віднесення вихідних даних до певного класу за допомогою виділення істотних ознак або властивостей, які характеризують ці дані, із загальної маси несуттєвих деталей.

Рішення завдання попередньої обробки зображення, виділення ознак і завдання отримання оптимального рішення і класифікації зазвичай пов'язане з необхідністю провести оцінку ряду параметрів. Це призводить до задачі оцінки параметрів. Крім того, очевидно, що виділення ознак може використовувати додаткову інформацію виходячи з природи класів.

Виділення завдань: пошук класів, опис класів, визначення найбільш ефективної системи ознак. Роль наявності або відсутності навчальної вибірки.

Виділення задачі автоматичної класифікації об'єктів (синоніми: багатовимірна класифікація, розпізнавання образів без вчителя, кластерний аналіз, таксономія).

Для збору даних у багатовимірному шкалі може бути використаний великий набір експериментальних методів.

Проте незалежно від цього, який метод обраний, дослідник повинен спочатку побудувати вибірку стимулів і набір ознак цих стимулів. Найкраще мати не менш як п'ять стимулів на кожну координатну вісь, виникнення якої очікується при аналізі.

Найчастіше застосовуваний експериментальний метод у тому, що випробувані дають прямі оцінки відмінностей всіх пар стимулів.

При зборі прямих оцінок відмінностей експериментатор може використовувати завдання з категоріальної оцінки, графічної оцінки, категоріального сортування або оцінки величини.

Експериментатори повинні давати завдання таким чином, щоб мінімізувати в оцінках випробуваних просторові та тимчасові ефекти. С. Розенберг розробили завдання щодо сортування, яке може бути використане для породження даних про спільну зустрічальність навіть тоді, коли число стимулів велике.

Дослідник повинен винайти спосіб побудови симетричної матриці близькості асиметричної матриці умовних ймовірностей.

Найчастіший захід різниці профілів — індекс відстані з формули (3.1). Однак до застосування цих заходів дослідник повинен вирішити, чи буде для підрахунку відмінності профілів оперувати початковими величинами, величинами, стандартизованими по стовпцях, величинами, центрованими по рядках, або величинами, стандартизованими по рядках. Це рішення залежить лише від змістовних міркувань, які у кожному додатку свої.

Ґрунтуючись на роботах Т. Юнга н А. Хаусхолдера [1938, 1941], У. Торгерсон [1952, 1958] запропонував один із перших алгоритмів МШ.

• Дж. Гоуер [1966, 1982] обговорив та розширив результати Торгерсона.

• Припущення Торгерсона значно жорсткіші проти сучасними методами. Тому підхід Торгерсонарідко використовують у початковому вигляді.

• У табл. представлено гіпотетичну матрицю відмінностей для шести спортивних ігор.

• Ця матриця відмінностей була побудована таким чином, щоб відобразити дві характеристики, за якими ці ігри різняться: швидкість спортивної гри та рівень контакту між гравцями.

Хокей та футбол – швидкі контактні ігри. Теніс та баскетбол – швидкі неконтактні ігри. Гольф та крокет – повільні неконтактні ігри.

• Дані із табл. будуть використані з метою показати, як можна застосовувати метод Торгерсон для побудови двовимірної конфігурації стимулів, що лежить в основі матриці даних. Це координатне застосування МШ.

• У моделі Торгерсона передбачається, що оцінки відмінностей дорівнюють відстаням у багатовимірному евклідовому просторі.

• Без втрати спільності можна припустити, що середнє значення координат стимулів кожної осі дорівнює нулю:

• Знайти матрицю X, яка задовольняє зазначеній умові, можна (якщо існує) з допомогою програми факторного аналізу методом головних компонент.

• Якщо розмір К не перевищує двох, то в додатках типу «стиснення даних» і «верифікація конфігурації» це питання є спірним.

• При такій невеликій кількості координатних осей важливі особливості конфігурації будуть видно просто при її розгляді, незалежно від повороту.

• Конфігурація на рис. ілюструє одну із можливих інтерпретаційних проблем. На одному кінці координатної осі I – повільні неконтактні ігри, на іншому – швидкі контактні ігри, а посередині – швидкі неконтактні ігри.

Шкала, представлена цією координатною віссю, неспроможна інтерпретуватися як швидкість кожної гри чи ступінь контакту. Це суміш того й іншого. Координатна вісь II також є сумішшю двох характеристик стимулів - швидкості та контакту. На її позитивному кінці – швидкі неконтактні ігри, а на іншому – повільні неконтактні та швидкі контактні.

• Вирішити питання про поворот конфігурації можна трьома способами. Якщо повернена конфігурація інтерпретована, то повертати її взагалі потрібно.

• На рис. ситуація не така. Отже, для отримання подання конфігурації, що інтерпретується, повинен бути застосований об'єктивний поворот або ручний поворот.

• Об'єктивні повороти

• Під об'єктивним поворотом розуміється математичний алгоритм знаходження інтерпретованого повороту конфігурації.

• Найбільш доступні об'єктивні повороти призначені для повороту отриманої в результаті факторного аналізу конфігурації тестів так, щоб повернена конфігурація задовольняла, наскільки це можливо, критерію простої структури

• Розв'язання задачі МШ задовольняє критерію простої структури, якщо кожен із стимулів має ненульове шкільне значення за однією характеристикою стимулів або, у крайньому випадку, за невеликою кількістю цих характеристик.

• Об'єктивні повороти

• У деяких додатках рішення, що добре інтерпретується, може дати об'єктивний поворот до такої простої структури, як варімакс або еквімакс.

• На рис. наведено поворот "варімакс" початкової матриці рішення.

• Поворот «варімакс» мало відрізняється від не повернутих рішень, тому проблеми інтерпретації тут ті самі. У цьому прикладі координатні осі рішення «варімакс» інтерпретувати не легше, ніж повернені осі.

• Об'єктивні повороти виконуються на комп'ютері.

• Ручні повороти виконуються людьми. Ручні повороти - це повороти, виконані дослідником і вибрані на основі зорового огляду не поверненої конфігурації.

• На практиці іноді можна, подивившись на конфігурацію, побачити, який поворот рішення дадуть координатні осі, що інтерпретуються.

• Якщо вісь I на рис. повернути на 45°, як показано на рис. , то всі швидкі спортивні ігри виявляться на її позитивному кінці, а всі повільні - на негативному. Можна сказати, що шкальні значення отриманої таким чином координатної осі відображатимуть швидкість спортивних ігор. Поворот на 45 ° осі II, як показано на рис. , призведе до осі, одному кінці якої будуть контактні ігри, але в іншому — неконтактні. Отримана координатна вісь відображатиме ступінь контакту у кожній грі.

• Можна швидко знайти кути між координатними осями I та II на рис. та координатними осями 1 та 2 на рис.

• Кут між осями I і 1-45 °, між осями II і 1 - 315 °, між осями I і 2-135 °, між осями II і 2 - 45 °.

• Відповідні косинуси дорівнюють 0,71, 0,71, -0,71 та 0,71. Об'єднання цих косінусів дає матрицю ортогонального перетворення:

• Помножуючи праворуч не повернуту матрицю шкальних значень X, зображену на рис. повернену матрицю координат.

• Координати можуть бути проінтерпретовані як ступінь контакту у спортивних іграх та їх швидкість.

• У координатних додатках пошук інтерпретованого повороту – важливий крок у процесі МШ. Можливі різноманітні підходи.

• Якщо початкове рішення інтерпретується, то досліднику взагалі не потрібний поворот. Якщо неповернене рішення погано інтерпретується, можна спробувати такий об'єктивний поворот, як варимакс чи эквимакс.

• Якщо ні неповернене рішення, ні об'єктивно повернене не призводять до інтерпретованих координат, дослідник може спробувати використовувати ручний поворот.

• РОЗМІРНІСТЬ

• Обговорення методу Торгерсона відбувається так, начебто кількість координатних осей К була відома.

• Однак на практиці вона не відома і має бути оцінена під час аналізу.

• Застосовуючи більшість методів МШ, користувач повинен отримати рішення в різних розмірностях і вибрати один із них, керуючись трьома критеріями:

• інтерпретованістю,

• відповідністю конфігурації даним

• відтворюваність.

• Алгоритм Торгерсона мінімізує наступну міру відповідності:

• Якщо стимул 1 – хокей, 2 – футбол, 3 – баскетбол, 4 – теніс, 5 – гольф, б – крокет, то для вирішення на рис. 4.1 друге власне значення одно:

• Корисним визначення розмірності може бути креслення, аналогічний рис.

• Вертикальна вісь має власні значення для не поверненого рішення, а горизонтальна вісь відповідає номерам координатних осей.

• Графік побудований за допомогою відкладення вгору від точки, що відповідає номеру осі, власного значення, пов'язаного із цією віссю. Наприклад, точка, що відповідає другій координатній осі, показує, що власне значення для другої осі неповірного рішення дорівнює 0,84.

• Якщо дані точно відповідають моделі рівняння (4.1), то графік повинен згладитись на (К+1) координатах так, як графік на рис. 4.4 згладжується на третій координаті.

• Іншими словами, на графіку має бути вигин на координаті, що перевищує правильне значення розмірності, К, на одиницю.

• У реальних даних, для яких немає точної відповідності моделі або великі помилки вимірювання та вибірки, вигин важко розрізнити. Справді, вигин на рис. 4. 4 розрізнити важко. У разі для правильного визначення розмірності графіка своїх значень то, можливо недостатньо. Потрібно розглядати і інтерпретованість, і відтворюваність.

• Відтворюваність може бути критерієм лише в тому випадку, якщо є не менше двох підвиборок.

• Основна ідея в тому, що слід зберегти в остаточному рішенні стільки координатних осей, скільки їх виявляться узгодженими в різних підвиборках.

• Якщо побудувати для кожної підвиборки своє рішення і у всіх підвиборках виявляться К. узгоджених координат, остаточне рішення має містити саме К координатних осей.

• Усі вибірки повинні бути взяті з однієї сукупності.

• Інтерпретація як критерій вимагає від дослідника деяких суб'єктивних оцінок. Однак при цьому рішення в більш високій розмірності воліє рішення в більш низькій розмірності, якщо існують важливі характеристики стимулів, що проявляються у рішенні вищої розмірності, але не виявляються у рішенні нижчої розмірності.

• І навпаки, рішення в більш низькій розмірності надається перевага, якщо немає таких суттєвих характеристик стимулів, які не виявляються у вирішенні низької розмірності.

• У прикладі в одномірному рішенні, що складається з зображеної на рис. 4.1 першої координатної осі, не можна розрізнити жодну з характеристик стимулів - ні ступінь контактності, ні швидкість.

• Обидві характеристики змішані єдиної координатної осі одномірного рішення.

• Тільки рішення, що містить дві координатні осі, може бути повернене таким чином, що кожна характеристика стимулів відповідатиме своїй осі на рис. 4.3.

• Оскільки двовимірне рішення інтерпретується легше, тут краще двомірне рішення.

**Тема 8. Проблема "стикування" змісту і формалізму при використанні алгоритмів класифікації**

1. Специфіка рішення соціологічних завдань побудови типології за допомогою методів автоматичної класифікації.
2. Сенс протиставлення термінів "класифікація" і "типологія". Підстава типології.
3. Роль апріорних уявлень дослідника про шуканих типах у виборі і реалізації алгоритму, інтерпретації результатів його застосування.
4. Виділення основних формальних елементів алгоритмів автоматичної класифікації, що вимагають стикування зі змістовними концепціями соціолога.

Специфіка рішення соціологічних завдань побудови типології за допомогою методів автоматичної класифікації. Сенс протиставлення термінів "класифікація" і "типологія". Підстава типології. Роль апріорних уявлень дослідника про шуканих типах у виборі і реалізації алгоритму, інтерпретації результатів його застосування. Виділення основних формальних елементів алгоритмів автоматичної класифікації, що вимагають стикування зі змістовними концепціями соціолога.

У процесі побудови моделей вивчення властивостей ми переконалися в тому, що в рамках кожної моделі потрібні певні типи інформації. Можна розглянути безліч підстав для виділення типів.

Поняття близькості між об'єктами є важливим поняттям методології аналізу соціологічної інформації. Чисто технічно виділення однотипних об'єктів зводиться до необхідності стиснення інформації. У змістовному аспекті це завдання вирішується в рамках однієї з мов аналізу, а саме типологічного аналізу в соціології.

Ці форми існування інформації виникають в багатьох галузях науки, які спираються на емпірію. Тому поза соціології існують наукові напрямки (аналіз часових рядів, методи дескриптивної статистики, багатовимірний статистичний аналіз і т. д.), де розроблені методи, прийоми, способи роботи з даними формами інформації. Зрозуміло, їх необхідно освоїти, але тільки в контексті змістовних завдань, які соціолог вирішує за допомогою цих методів.

Розглянутий вище тип інформації, з точки зору соціолога-користувача, володіє двома недоліками: можливими неповнотою і недостовірністю. Перше полягає в тому, що вона може не містити інформації, що цікавить соціолога. Друге означає наступне. Наприклад, відомий факт, що в процесі перепису населення жінки занижували свій вік. Це призводить до неможливості правильного прогнозу частки населення пенсійного віку в певні роки. Відомо також заниження показників дитячої смертності в роки, коли ця статистика була закритою.

ІНТЕРПРЕТАЦІЯ

• Інтерпретованість обговорювалася вище, коли йшлося про вибір розмірності.

• Інтерпретація рішень залежить від вибору істотних характеристик стимулів. Такі характеристики це зазвичай впорядкування або угруповання стимулів.

• Суттєво важлива група стимулів — це набір стимулів, що групуються разом, в одній області багатовимірного простору рішення, які мають будь-яку загальну ознаку.

Наприклад, у дослідженні професій торгові професії можуть розташовуватися разом, утворюючи розумне угруповання. При дослідженні популярних журналів можуть групуватися разом журнали для жінок

• Істотне впорядкування стимулів – це впорядкування, що відповідає порядку стимулів за їхньою важливою характеристикою.

Наприклад, упорядкування стимулів координатної осі I на рис. 4.3 відповідає їх упорядкування за ступенем контакту у грі. Впорядкування по осі II відповідає впорядкуванню ігор за швидкістю.

• Обидві ці координатні осі є суттєвими впорядкуваннями, оскільки вони відповідають важливим характеристикам стимулів — швидкості та ступеня контакту в грі.

• В ідеалі координатні осі мають бути повернені таким чином, щоб кожна з них представляла одне із суттєвих впорядкувань.

• Інтерпретація рішення включає ідентифікацію важливих угруповань та впорядкування стимулів. Для угруповань необхідно ідентифікувати ті риси, які є спільними

всім об'єктів кожного кластера. Для впорядкування потрібно ідентифікувати відповідні ознаки. Один із способів інтерпретації рішення — простий розгляд конфігурації.

• Р. Сміт та А. Зігель [1967] використовували МШ при визначенні координатних осей для обов'язків керівників цивільної оборони (ГО).

• У три послідовні етапи вони ідентифікували 34 обов'язки, які, на їхню думку, представляють усі обов'язки керівників ДО.

• Тридцять п'ять керівних працівників ГО оцінювали різницю кожної пари обов'язків за шкалою з 11 пунктів. Потім для отримання матриці відмінностей оцінки відмінностей 35 випробовуваних були об'єднані.

• Для отримання чотиривимірного рішення використовувався алгоритм Торгерсона. Сміт та Зігель використовували об'єктивний поворот еквімакс

• Для кожної із чотирьох координатних осей у табл. 4.2 наведено обов'язки, що мають найвищі позитивні значення та найнижчі негативні значення.

• Автори резюмують їхню інтерпретацію координатних осей у присвоєних назвах:

 внутрішня підтримка системи - зовнішня експлуатація системи (вісь I),

 регулярне планування - планування на надзвичайний випадок (вісь II),

 використання ресурсів - оцінка ресурсів (вісь III),

 інтеграція системи надзвичайних обставин (вісь IV)

• Сміт та Зігель пропонують використовувати отримані координатні осі як основу для розробки одновимірних шкал оцінок працівників ГО.

Наприклад, може виникнути бажання створити шкалу оцінок працівників ГО, відповідну осі III, де завдання оцінки ресурсів з'являються одному полюсі.

• Автори пропонують навіть планувати за допомогою отриманих осей відбір та навчання працівників ГО.

ІНШІ МЕТРИЧНІ МОДЕЛІ

• Теорія максимальної правдоподібності, що лежить в основі цих алгоритмів, робить можливим розробити міру відповідності, яка при нульовій гіпотезі, що представляється шкальною моделлю, розподілена приблизно до розподілу хі-квадрат. Ранні версії алгоритму Рамсея вимагали стільки машинного часу, що могли використовуватися лише невеликих масивів даних. Якщо подолати обчислювальні труднощі, то підхід, заснований на методі максимальної правдоподібності, дозволить дослідникові перевірити відповідність своїх даних моделі строго, ніж це можливо за інших моделей.

• Рішення (конфігурація) може залишитися неповерненим, може бути повернуто вручну або за допомогою будь-якого об'єктивного алгоритму, такого як варімакс або еквімакс.

• З цих трьох способів повороту краще той, який дає найбільш інтерпретовані напрямки. Інтерпретація рішення включає ідентифікацію угруповань стимулів або впорядкування стимулів, що відповідають їх суттєвим характеристикам.

• Сміт та Зігель використовували алгоритм Торгерсона для побудови координатних осей обов'язків. Вони дійшли висновку, що отримані координатні осі можуть бути основою створення одномірних шкал оцінок працівників ГО і програм їх навчання.

**Тема 9. Функції відстані між об'єктами**

1. Аксіоматичне визначення функції відстані і ролі цієї функції в соціології.
2. Приклади непридатності евклідової відстані з точки зору апріорного змістовного розуміння типів об'єктів.
3. Можливість використання евклідової відстані в розглянутих прикладах за рахунок зміни ознакового простору.
4. Розгляд факту як однієї з реалізацій загального принципу органічного зв'язку між виміром та аналізом зібраних з його допомогою даних.
5. Функції відстані, відмінні від евклідова: зважене евклидово, сіті-блок, Махаланобіса, Хеммінгово.

Аксіоматичне визначення функції відстані і ролі цієї функції в соціології. Приклади непридатності евклідової відстані з точки зору апріорного змістовного розуміння типів об'єктів.

Можливість використання евклідової відстані в розглянутих прикладах за рахунок зміни ознакового простору.

Розгляд цього факту як однієї з реалізацій загального принципу органічного зв'язку між виміром та аналізом зібраних з його допомогою даних.

У сучасних умовах науково-технічного розвитку в усіх сферах діяльності людини стало аксіомою прийняття рішення на основі аналізу даних. Способи, методи отримання інформації з даних та вироблення нових знань є супроводом у системах підтримки ухвалення управлінських рішень в управлінні об’єктами різної природи. Метою аналізу даних є вивчення властивостей об’єктів, явищ та процесів, отримання нових знань про них для більшого підпорядкування. Сучасний аналіз даних обумовлюється способами отримання величин, методами їх обробки й залежить від розвитку математичних методів і моделювання. Дана ситуація є типовою для всіх сфер діяльності людини, наприклад, директор фірми на основі даних про діяльність підрозділів намагається скласти об’єктивне уявлення про їх функціонування; працівники економічного управління намагаються вивчити основні тенденції

економічного і соціального розвитку регіону на основі системи показників протягом встановленого періоду; спеціалістам науково-експертного управління країни потрібно вивчити й достовірно порівняти економічний та соціальний стан областей. Перелік прикладів можна продовжити до нескінченності, але всі вони потребують використання методів аналізу даних для впорядкування наявної інформації, подання її в лаконічній, узагальненій, стислій, очевидній формі, яка полегшує процедуру формування управлінського рішення за виявленими тенденціями, закономірностями, вилученими новими знаннями Функції відстані, відмінні від евклідова: зважене евклидово, сіті-блок, Махаланобіса, Хеммінгово.

### Тема 10. Основні види процедур класифікації. Відстані між класами

1. Виділення ієрархічних і неієрархічні алгоритмів класифікації.
2. Агломеративні та дівізімні алгоритми.
3. Оптимізація розбиття як один з основних елементів формалізму в неієрархічні алгоритмах класифікації.
4. Способи вимірювання сумарних оцінок близькості один до одного об'єктів усередині класів.
5. Приклади соціологічних задач, для яких змістовно адекватні різні способи вимірювання відстаней між класами.

Актуальність дослідження сутності та методів багатовимірного аналізу соціологічної інформації обумовлена специфікою соціальної реальності, що завжди уявляється як складний, багатогранний та багатозначний феномен, який інтегрує багатовимірність суспільства з багатовимірністю внутрішнього світу окремої людини. Соціологи, вивчаючи соціальну реальність, стикаються з необхідністю вибору адекватних підходів та методів, здатних охопити всі аспекти досліджуваних соціальних явищ, враховуючи їх цілісність та взаємозалежність. Як відомо, на теоретичному рівні вирішення цього завдання здійснювалося шляхом розробки різноманітних теоретичних підходів, які склали основу поліпарадигмальності сучасної соціології. Виділення ієрархічних і неієрархічні алгоритмів класифікації. Багатовимірний статистичний аналіз (у широкому значенні) - розділ математичної статистики, що поєднує методи вивчення даних, які характеризують багатовимірні об'єкти. Багатовимірний статистичний аналіз (у вузькому значенні) поєднує ті багатовимірні статистичні методи, які засновані на припущенні, що результати окремих спостережень незалежні й підлеглі багатовимірному нормальному розподілу. Звичайно саме до цієї частини математичної статистики застосовують термін "багатовимірний статистичний аналіз".

Агломеративні та дівізімні алгоритми. Причини необхідності розгляду відстаней між класами в ієрархічних процедурах. Алгоритм найближчого сусіда як приклад способу класифікації, що використовує такі відстані.

Оптимізація розбиття в сенсі максимізації заздалегідь обраного функціоналу якості як один з основних елементів формалізму в неієрархічні алгоритмах класифікації. Основний змістовний сенс такої оптимізації - прагнення до того, щоб усередині класів об'єкти були якомога ближчими один до одного, а класи були б якомога далі один від одного. Сенс вимірювання близькості між класами в таких випадках. Способи вимірювання сумарних оцінок близькості один до одного об'єктів усередині класів. Приклади соціологічних задач, для яких змістовно адекватні різні способи вимірювання відстаней між класами.

**Тема 11. Гіпотези про розташування об'єктів у ознаковому просторі**

1. Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації.
2. Основні види гіпотез: компактності, зв'язності (безперервності), унімодального розподілу.
3. Приклади соціологічних завдань побудови типології
4. Загальне уявлення про розмиті класифікації.
5. Доцільність комплексного використання декількох алгоритмів класифікації в соціологічних завданнях побудови типології.
6. Змістовні уявлення соціолога про типи та умови вибору кроку розбиття при інтерпретації результатів. Коригування результатів класифікації з метою забезпечення відповідності класифікації і типології.

Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації. Обумовленість цих гіпотез апріорними уявленнями дослідника про типи об'єктів. Основні види гіпотез: компактності, зв'язності (безперервності), унімодального розподілу.

Факторний аналіз найбільш яскраво відображує риси багатомірного аналізу в частині дослідження зв'язку між ознаками. Кластерний аналіз ці риси відображує з боку класифікації об'єктів. Сіизіег (англ.)- нагромадження груп елементів, які характеризуються якою - небудь загальною властивістю. Суть його зводиться до групування (кластеризації) сукупності з різноманітними ознаками з метою одержання однорідних груп - кластерів. При цьому межі таких груп наперед не завдані, а кількість їх може бути або завдано, або ні. Одержані в результаті розмежування групи називаються кластерами, а методи їх знаходження - кластер-аналізом. У кластерному аналізі ознаки об'єднуються в один кількісний показник схожесті (несхожесті) групуючих об'єктів.

Приклади соціологічних завдань побудови типології, для яких була б розумна кожна гіпотеза. Приклади алгоритмів, що шукають закономірності розташування точок у ознаковому просторі, що відповідають кожній з гіпотез: алгоритм Форель (гіпотеза компактності), алгоритм найближчого сусіда (гіпотеза зв'язності), алгоритм, заснований на виділенні локальних максимумів функції приналежності (гіпотеза унімодального розподілу ).

Загальне уявлення про розмиті класифікації. Роль функції належності у відповідних алгоритмах. Доцільність комплексного використання декількох алгоритмів класифікації в соціологічних завданнях побудови типології.

Змістовні уявлення соціолога про типи та умови вибору кроку розбиття при інтерпретації результатів. Коригування результатів класифікації з метою забезпечення відповідності класифікації і типології.

**Тема 12. Поняття інтерпретації вихідних даних і основні методологічні принципи використання методів аналізу даних в соціології**

1. Інтерпретація вихідних даних як одне з основних ланок "стикування" соціології і математики.
2. Основні фактори, що визначають інтерпретацію вихідних даних: апріорні уявлення дослідника про спосіб породження цих даних
3. Виділення методологічних принципів.

Можна було б говорити ще про цілу низку подібних вимог, що носять більш приватний характер: необхідність виконання деяких принципів вимірювання цікавлять соціолога показників; забезпечення певної однорідності тієї сукупності об'єктів, на якій "діє" наша передбачувана закономірність; дотримання деяких принципів інтерпретації результатів застосування методу; виконання певних правил комплексного використання цілої серії методів при вирішенні практично будь-якої соціологічної завдання і т.д.

Розкриття кожного з названих принципів вимагає серйозного розгляду. Всі вони багатоаспектний, мають складну структуру. Їх практична реалізація вимагає досить глибокого аналізу концептуальних уявлень соціолога про вивчається явище, для чого потрібно чітке формулювання самих цих уявлень.

Інтерпретація вихідних даних як одне з основних ланок "стикування" соціології і математики. Основні фактори, що визначають інтерпретацію вихідних даних: апріорні уявлення дослідника про спосіб породження цих даних (у тому числі – про моделі сприйняття респондентами пропонованих ним питань, об'єктів, про ймовірнісну природу даних і т. д.); мета дослідження; концептуальні уявлення соціолога про досліджуване явище; характер

моделі явища, "закладеної" в математичному методі, використання якого планується; розгляд спостережуваних змінних як непрямих показників латентних факторів, насправді цікавлять дослідника і т. п.

Виділення методологічних принципів, дотримання яких є необхідним для того, щоб аналіз соціологічних даних був ефективний, не відводив соціолога в сторону від реальності: забезпечення певної однорідності вихідних даних; облік моделі, "закладеної" в кожному методі аналізу даних, при виборі алгоритму аналізу, два основні принципи інтерпретації результатів аналізу: необхідність її узгодження з інтерпретацією вихідних даних і заповнення при її здійсненні тих втрат, які мали місце при переході до формалізму; необхідність комплексного використання декількох методів для вирішення одного завдання і т. д.

**Тема 13. Великі дані у соціології: нові дані, нова соціологія?**

1. Визначення даних. Філософський, юридичний підходи й життєвий цикл даних.
2. Поняття метаданих. Життєвий цикл метаданих
3. Оцінка вимог та аналіз контенту
4. Специфікація системних вимог. Система метаданих

Історію соціальних наук можна як зміну етапів, пов'язаних з характером домінуючих даних (Mohretal., 2013: 676). Все минуле століття нашими джерелами даних служили опитування, інтерв'ю та спостереження. Зараз розпочався наступний етап, у якому вирішальну роль відіграють нові технології виробництва та збору великих даних про ті аспекти поведінки людини, які раніше не піддавалися спостереженню. Великі дані з'являються не в результаті опитувань або інтерв'ю, їх створення опосередковано технологіями:

мобільні телефони, електронна пошта, сервіси онлайн-банку, транзакції кредитних карток, переміщення по сайтам, зчитування бар-кодів, соціальні мережі тощо. Новий лейбл «обчислювальна соціальна наука» (computationalsocialscience) все частіше використовується для позначення дослідницького поля, в рамках якого поведінка людини аналізується за допомогою нових способів виробництва, обробки та методів аналізу даних (Lazeretal., 2009). Навколо нових ідей розвивається інфраструктура: запущено спеціальне фінансування, відкриваються нові магістерські програми, створюються журнали та дослідницькі центри.

У цьому огляді зроблено спробу відповісти питанням, які зміни привнесли нові дані у соціологію. Якщо піти простим шляхом, то можна розглянути, як нові можливості використовуються в соціології.

Інший шлях полягає в тому, щоб звернутися до змін, що стосуються всього дисциплінарний проект соціології У цьому випадку ми не говоримо про ті дослідницькі галузі, для яких великі дані відкрили друге дихання, але спробуємо передбачити, чи зміниться сама дисципліна. Соціологи пропонують замінити традиційну соціологію на доказову соціальну науку, яка відрізняється від звичного для мейнстриму стилю з послідовним рухом від гіпотез до збирання та аналізу даних.

Спочатку ми звернемося до дискусії про те, що складає концептуальні особливості великих даних, які дозволяють називати їх не так великими. скільки новими даними. У наступній частині йтиметься про ті області соціології, у яких відзначено помітний інтерес до нових даних. Так, великі дані виявилися важливими для соціології у двох відносинах. По перше, вони надали можливість вивчати соціальну поведінку, доступ до якої раніше був обмежений (McFarland, Lewis, Goldberg, 2015). Робота з онлайн-даними дозволила просунутися у вирішенні теоретичної проблеми, пов'язаної з визначенням природи соціального впливу - передачі через соціальні зв'язку патернів поведінки, установок, хвороб чи навіть емоцій. По-друге, запозичення інструментів комп'ютерної науки змінило спосіб аналізу великих неструктурованих масивів тексту, що особливо важливо для тих наукових областей, де досліджується символічне провадження. У заключній частині йтиметься про зміни в дослідному стилі соціології.

Наш огляд має свої обмеження. Він присвячений можливостям застосування великих даних саме в соціології і не торкається інших дисциплін. Ми також не стосуємося обмежень, пов'язаних із застосуванням великих даних, у тому числі етичного характеру.

**Про природу великих даних**

Соціальні науки далеко не одразу піддалися чарівності нових можливостей, перші наукові статті з'являються лише у 2009 році. На той час ера великих даних вже було проголошено масовими виданнями, що пов'язані з появою нової діяльності — аналітики даних у комерційному секторі. Кількість статей у масовій періодиці досі помітно перевищує кількість статей у наукових журналах. Великі дані виявилися корисними в комерційному секторі для покращення робочих операцій та вилучення більшого прибутку, оскільки надали можливість передбачати поведінку людей на основі даних, що вже існують. Аналізуючи те, що цікавить людей у цю хвилину — на яких сайтах вони проводять час і які запити надсилають до пошукових систем, можна передбачити, чого вони захочуть у найближчому майбутньому.

Сучасну історію великих даних іноді починають зі слів дослідників NASA, які в 1997 зіткнулися з тим, що їх комп'ютери не справляються з обсягом даних. Таким чином, спочатку акцент робився на обсязі: великі дані - це дані, з якими не справляється Excel на простому

комп'ютер. Однак якщо рахувати обсяг за головний параметр, то доведеться визнати відносний характер великих даних, оскільки можливості обробки великих масивів інформації постійно удосконалюються (Austin, Fred, 2016).

Надалі йшлося про три характеристики — розмір, швидкість накопичення та різноманітність (volume, velocity, variety). Згідно з Р. Кітчином, великі дані відрізняються більшим обсягом; високою швидкістю накопичення (вони створюються тут і зараз і їх обсяг може збільшуватись кожну секунду); різноманіттям форми; вичерпним характером (найчастіше представляють всю сукупність); високою дискретністю (що дозволяє дробити дані на окремі групи та легко їх ідентифікувати); можливістю прив'язки до інших типів даних; гнучкістю (додавати нову інформацію та розширювати обсяг) (Kitchin, 2014: 2).

Незважаючи на деякі спроби описати ключові характеристики великих даних, ми можемо говорити швидше про лейбл, який схоплює різні дані в одному найменуванні (Kitchin, McArdle, 2016). Цей термін втратив концептуальну ясність, що добре показано на прикладі аналізу Р. Кітчіном та Дж. Макардлом 26 наборів даних, використаних у наукових дослідженнях.

Дані під одним лейблом мають як загальні, і відмінні друг від друга характеристики. Більш того, дослідники не виявили жодного набору даних . Основні критичні аргументи можна знайти у: Boyd, Crawford, 2012; Zwitter, 2014; Iliadis, Russo, 2016, який описувався через всі сім ключових характеристик. Вони визначили лише дві ключові риси, яким відповідали всі 26 досліджень.

Це швидкість накопичення і всеосяжний охоплення (вся реальність об'єктів цього типу, n = all). Під ці характеристики підпадають головним чином онлайн-дані чи дані, що створюються за рахунок електронних технологій. Всі данні, які аналізували Кітчин та Макардл, так чи інакше, передбачали використання електронних засобів. Є лише один виняток із цього списку — адміністративні дані, що генеруються державними відомствами.

При всій відмінності від онлайн-даних їх, проте, можна вважати великими даними через те, що вони зазвичай охоплюють всю популяцію і виробляються в реальному часі, хоча при цьому доступ до них може припускати тимчасовий лаг (Connellyetal., 2016).

Отже, бачимо, що розмір перестав бути необхідною умовою визначення суті великих даних. Справді, і раніше існували дані, які були досить великими, наприклад, дані перепису, з одного боку, містили інформацію про тисячі одиниць, але з іншого — не були гнучкими, їх важко було доповнити іншими даними, і були потрібні спеціальні зусилля по їх генерації (Kitchin, 2014). Дослідники вважають, що революція у великих даних відбулася не тому, що тепер можна мати справу з даними великого обсягу, головне, що дані створюються не для цілей дослідження (Connellyetal., 2016: 2). Раніше вони збиралися на запит дослідника,

найчастіше заздалегідь певної процедури та відповідно до дослідницьких припущень чи гіпотезами. Зараз дані виробляються самими користувачами: люди пишуть пости, ставлять лайки, завантажують фотографії і роблять покупки, у свою чергу, державні відомства стають власниками даних про різні галузі — освіту, медицину, кримінологію (Волков, Скугаревський, Титаєв, 2016). Збулася мрія соціолога — дістатися слідів, що залишаються від дій людей, незалежно від намірів дослідників. Причому «більше немає необхідності вибирати між кількістю одиниць у наших даних та кількістю інформації про них… Детальна інформація та розуміння, яке раніше можна було отримати тільки про небагатьох, зараз доступні для великої кількості людей» (Manovich, 2011).

Отже, створення нових даних майже не пов'язане з намірами вчених провести дослідження. Вони поєднують два аспекти, які раніше практично не зустрічалися разом, це масштабні дані про поведінку людей на мікрорівні. Що нового це означає для соціології?

**Нові дані: соціальний вплив**

Маніфестом «нової науки» можна вважати статтю «Обчислювальна соціальна наука», яка з'явилася у «Science» у 2009 році (Lazeretal., 2009). Автори статті не раз виступали в ролі ключових спікерів великих профільних конференцій, вони керують центрами та інститутами, результати їх досліджень з'являються в престижних «Science» та «Nature». Вихідна теза: великі дані не можуть не змінити соціальну науку, оскільки даних такого масштабу на рівні тонких взаємодій раніше не було. Ідея викликала інтерес з боку представників різних дисциплін. Інформація про посилання на статтю дозволяє зробити висновок про те, що увагу до неї забезпечують вчені, які публікуються в одних із найпрестижніших журналів: "PlosOne", "PNAS", "ScientificReports", "Science" (разом - третина всіх статей). При цьому особливий інтерес можна відзначити саме у соціологів — на це вказує список десяти найцитованіших журналів у статтях, що посилалися на «Обчислювальну соціальну науку».

Цей список багато в чому складається з міждисциплінарних журналів, проте серед дисциплінарних на перших місцях знаходяться соціологічні видання. Насамперед соціологи бачать переваги у можливості просунутися за рахунок нових даних та способів їх аналізу. Гері Кінг вважає, що найбільший результат у соціальній науці можливий, коли є три умови: інноваційні статистичні методи, нова комп'ютерна наука та оригінальні теорії окремих галузей знання (King, 2013). У цьому сенсі соціальні науки повинні займатися всім тим самим, але з кращими методами і кращими даними, які дозволяють подолати недоліки колишніх даних — їх штучні умови створення, ретроспективний характер та статичність інформації, що збирається (Golder, Macy, 2014).

Замість того, щоб намагатися кожні два роки отримати думки про політику у кількох тисяч активістів шляхом штучно створеної ситуації розмови у вигляді опитувального інтерв'ю

ми можемо використовувати нові методи і отримати десятки мільйонів політичних думок, які з'являються щодня у блогах. Також як замість того, щоб вивчати вплив контексту на взаємодії людей, запитуючи респондентів про їх останні контакти, ми можемо зібрати інформацію за тривалий проміжок часу їх телефонних дзвінках, листах та повідомленнях. За відсутності офіційної статистики ми можемо судити про економічний розвиток і зростання населення, ґрунтуючись на інформації зі знімків супутника про освітлення, розташування доріг та інших об'єктів інфраструктури. (King, 2009: 92).

Наприклад, А. Пентланд очолює в MIT лабораторію з вивчення динаміки поведінки людини. А.-Л. Барабаші, відомий дослідник мереж та творець окремого напряму науки про мережі, очолює Центр досліджень складних мереж. Широкій публіці має бути відомий Н. Крістакіс, автор бестселера «Пов'язані однією мережею. Як на нас впливають люди, яких ми ніколи не бачили» (у співавторстві з Дж. Фоулер). У ЄліКрістакіс очолює Лабораторію з досліджень природи людини, а також Інститут досліджень мереж. Слід зазначити, що у цій зірковій компанії ми бачимо мікс соціальних вчених та тих, хто не отримував ступеня в соціології чи політичній науці. З 15 авторів статті половина має ступеня з соціальних наук, інші писали дисертації у сфері фізики та комп'ютерних наук.

Нові дані тим самим розширюють простір і дають нові можливості для розвитку звичних напрямків досліджень, особливо тих, які скористатимуться онлайн-даними. Збір даних щодо поведінки людей у різних контекстах складний, вимагає ресурсів, а деяких випадках проблема доступу така серйозна, що деякі дослідницькі питання залишаються незаданими. Онлайн-дані надають інформацію про поведінку людей у реальному часі, фіксуючи автоматично, хто, де та з ким зараз взаємодіє; при цьому мінімізується вплив дослідника при самому виробництві даних, адже вони існують незалежно від того, чи буде їх аналізувати чи ні (Golder, Macy, 2014). Є думка, що онлайн-дані змінили соціальні науки, так само як свого часу електронний мікроскоп чи МРТ змінили природничі науки — нові інструменти дозволили спостерігати за онлайн-активністю, і саме це справляє трансформуючий ефект на соціальну науку (Golder, Macy, 2014).

Головний результат для теоретичної соціології, мабуть, полягає у появі можливості використовувати онлайн-дані вивчення соціального «зараження» — передачі через соціальні зв'язку патернів поведінки, установок, хвороб чи навіть емоцій (Christakis, Fowler, 2013). Використання епідеміологічної метафори вимагало уточнення, який механізм задіяний при "зараженні". Справді, якісь речі — мікроби чи інформація можуть передаватися від людини до людини без особливої участі мереж, а через швидкоплинні контакти. Однак для поширення поведінки може бути необхідна мережа наявність певних структурних характеристик (Smith, Christakis, 2008: 412). Виникла дискусія про те, який тип зв'язків необхідний поширення

феноменів по мережах. Класичні роботи показали важливість слабких зв'язків або структурних дірок у мережах, які потрібні для переміщення та матеріальних ресурсів, та інформації (Granovetter, 1973; Burt, 2004). Слабкі зв'язки мають ту особливість, що далеко простягаються, а отже, можуть досягти більшої кількості людей, на відміну від сильних зв'язків, які мають тенденцію до кластеризації. Однак залишається питання: чи достатньо контакту через слабкі зв'язки, щоб відбулася успішна передача? Інша гіпотеза полягала в тому, що соціальна поведінка має на увазі складний вплив: людям зазвичай потрібно вступити в контакт із кількома джерелами «інфекції», перш ніж вони пішли: державна статистика, опитування населення та інтерв'ю з представниками влади. Кожен з цих джерел мав свої серйозні недоліки, щоб можна було покладатися на них щодо цензури. Дослідження Кінга будувалося на аналізі цензури онлайн-записів. Збір даних передбачав вилучення з численних сайтів записів, доки вони не були прочитані цензорами і видалені (всього було зібрано 3674698 записів). Надалі відслідковувалося, чи було зроблено втручання цензора (це сталося 13%). Категорії, які цензурувались, — події, що стосуються колективної дії, критики цензорів та порнографії, тоді як категорії, в яких обговорювалися рішення уряду, не проходили через цензуру. Державні лідери навряд чи раді критичним зауваженням, проте це не турбує їх настільки, щоб задіяти цензуру видалення критичних записів. Що їх справді турбує, то це події, які можуть сприяти згуртуванню людей (King, Roberts, 2013) почуваються готовими запозичити поведінкові зразки (Centola, 2011; Centola, 2010). Тоді успіх у соціальному «зараженні» забезпечується скоріше сильними зв'язками, які «надлишкові» за своїм характером.

Можливість аналізувати великі мережі дозволила просунутися у цьому напрямі. Шукати відповіді колишніми способами було надто дорого. Опитувальні інструменти будуються на вибіркових механізмах, цим спочатку зв'язку між людьми обмежуються дизайном дослідження – вже не можна поставитися питанням, як організовано вплив у широких мережах (Golder, Macy, 2014). Соціальні вчені у мережевому аналізі частіше використовували етнографічні методи, збирали інформацію для аналізу статичних мереж невеликого масштабу (McFarland, Lewis, Goldberg, 2015). Якщо і виходило зібрати дані про зв'язки, щоб виміряти соціальний вплив, то витрати не дозволяли включити в аналіз значну кількість випадків. Елізабет Ботт (Bott, 1955) проводила вечори за довгими бесідами, щоб зафіксувати інтенсивність зв'язків та взаємодій, коли вивчала лондонські сім'ї (загалом у дослідженні брали участь 18 сімей). У свою чергу, Б. Веллману (Wellman, 1979) вдалося зібрати набагато більше мережевих даних про канадських робітників, проте опитування давало можливість поставити досить короткий перелік питань. У підручниках з мережевого

аналізу є рекомендації, що респондентів варто просити надати інформацію про їх зв'язках не більше ніж з п'ятьма людьми.

### Тема 14. Великі дані. Системи керування великими даними

1. Розподілені файлові системи
2. Розподілені фреймворки
3. Бенчмаркінг
4. Серверне програмування
5. Планування
6. Системи розгортання

Великі дані можуть бути різних типів. Інформацію, отриману в результаті обліку або вимірювання будь-яких об'єктів або параметрів, називають майстер-даними (MasterData). Наприклад, облік кількості, виміри координат і швидкостей конкретних молекул - це майстер-дані.

Транзакційні дані (в англомовній літературі застосовуються терміни TransactionalData, ApplicationSpecificData, OperationalData) – це дані, що відображають результат виконання будь-яких операцій. Транзакційні дані описують взаємодія об'єктів один з одним або з навколишнім світом, які можна отримати за допомогою обробки майстер-даних.

Ретроспективні дані (Historicaldata) – це дані, забезпечені позначки часу.

Посилальні дані (довідники, НСІ, нормативно-посилальна інформація,

ReferenceData, LookupData, Dictionaries) – це базові незмінні дані, заздалегідь відомі із зовнішніх джерел, такі як нормативи, скорочення, акроніми, словники, стандарти.

Формат даних. Структуровані дані мають заздалегідь визначений формат. Напівструктуровані або слабоструктуровані дані - це дані, які часто зібрані з різних джерел.

Дослідники змушені були будувати гіпотези, які враховували соціальний вплив через сильні зв'язки окремих людей, що не давало змоги схоплювати структурні характеристики мереж (інформацію про те, як виглядає вся мережу контактів людини). Зараз же через те, що взаємодії залишають свій онлайн-слід, є можливість зібрати для великої кількості людей інформацію про їх телефонні дзвінки, електронні листи, смс-повідомлення, адресних книгах та онлайн-взаємодіях у соціальних мережах (King, 2009). Іншими словами, з'явилися інструменти, які настільки полегшили збирання даних про відносини, що свого роду відкриття науки про мережі сталося заново.

Поява соціальних мереж та реєстрації онлайн-поведінки уможливили проводити онлайн-експерименти, які дозволяють контролювати Більшість умов, а також оцінити ефект

втручання на великих вибірках індивідів (McFarland, Lewis, Goldberg, 2015). Один з найвидатніших прикладів заснований на даних понад 60 мільйонів користувачів Фейсбуку (Bondetal., 2012). Експеримент тестував можливості соціального впливу під час політичної мобілізації на прикладі виборів до Конгресу. В експерименті 2010 року брали участь усі американські користувачі Фейсбуку з 18 років, які були поділені на три групи. Першій групі (60 055 176) у стрічці новин показали «соціальне повідомлення»: у ньому містився заклик проголосувати, інформація про місце голосування, а також показувалися профілі людей з-поміж друзів користувача, які вже проголосували. У стрічці другої групи (611044) з'явилося інформаційне повідомлення, в якому також була кнопка, на яку можна було натиснути і цим показати друзям, що ти проголосував; тут також присутнє посилання про місце голосування, проте соціальна складова повідомлення не було. Третя група – контрольна – взагалі не отримала жодних повідомлень. Дії, що потім аналізувалися: натискання ярлика «I vote», перехід за інформаційним посиланням та голосування на виборах.

Згідно з результатами, ті, хто отримав соціальне повідомлення, частіше натискали на кнопку про те, що вони проголосували, ніж ті, хто отримав лише інформаційне повідомлення. Цей результат повторився і на даних про реальне голосування, тобто люди з першої групи частіше голосували, ніж користувачі з інших груп експерименту. Між контрольною групою та групою лише з інформаційним повідомленням взагалі не було статистично важливих відмінностей. Це говорить про те, що тільки інформація не особливо змінює поведінку людей, тоді як соціальний тиск виявляється дієвим, змінюючи поведінку про політичне самовираження (розповісти друзям про те, що я проголосував) та участь у реальному голосуванні.

В експерименті також спробували перевірити, наскільки вплив залежить від сили зв'язку (частота взаємодій як обміну повідомленнями). Виявилося, що близькі друзі людини, яка сама натиснула на кнопку, також частіше самі натискали на кнопку I vote і частіше голосували, ніж близькі друзі тих, хто був у контрольній групі. Інші друзі, які формували з користувачем слабкі зв'язки, виявилися незачепленими впливом — вони не стали частіше повідомляти про те, що проголосували, так само як вони не стали частіше справді голосувати. Результати підтверджуються в інших онлайн-експериментах (Covieloetal., 2014).

В огляді «AnnualReviewofSociology» підбиваються підсумки про внесок даних про онлайн-поведінці у розвиток соціальних наук (Golder, Macy, 2014). Величезний сплеск інтересу часто обертається дослідженнями низької якості. Багато статей повторюють ідеї попередніх авторів, але більш масштабних даних, причому без будь-якої відсилання до класичних робіт. Створюється враження, що мережі дають такий інструмент, що дозволяє

побудувати графи майже на будь-яких даних, що позбавляє його свідомості (boyd, Crawford, 2012).

Водночас саме аналіз великих мереж вбудовується у класичні соціологічні сюжети про природу соціального впливу, які при цьому виконуються на високому методологічному рівні. Цей напрямок досліджень з'єднує всі три складові, про які писав Кінг, — інноваційні статистичні методи, нова комп'ютерна наука та оригінальні теорії окремих областей знання.

Далі ми побачимо, що нові можливості для соціології з'являються не лише з доступом до даних, які раніше були надто дорогими чи зовсім були відсутні, але з розвитком інструментів роботи з даними, доступ до яких існував завжди.

**Нові методи аналізу текстових даних**

Нові дані перевершують старі у своєму обсязі, різноманітності та глибині, але зазвичай існують зовсім у тому вигляді, у якому готові до аналізу. Перетворення сирих даних на необхідний дослідників формат вимагає спеціальних компетенцій у сфері комп'ютерної науки. Дослідники перераховують цілий арсенал методів: математичне та статистичне моделювання; динамічний аналіз мереж; автоматичне генерування гіпотез; методи інтеграції мультимодальних даних; можливості обробки природної мови та машинне навчання (Golder, Macy, 2014). Соціологам потрібні люди, які вміють програмувати не тільки для того, щоб отримувати дані, а й для того, щоб їх аналізувати. Вирішення цих завдань призвело до успіхів — автори пишуть про чотири прориви в аналізі даних. Перший пов'язаний з великими масивами текстових даних та відсилає до галузі обчислювальної лінгвістики, другий розвиває мережевий аналіз, третій спирається на досягнення машинного навчання, нарешті, четвертий використовує можливості онлайн-експериментів (McFarland, Lewis, Goldberg, 2015). З цього списку особливої уваги заслуговують інструменти в галузі обчислювальної лінгвістики. Поява тематичного моделювання описується як крок революційного значення, який даний момент поки що не оцінений соціологами належним чином (Evans, Aceves, 2016). Головні сфери застосування — це соціологія науки та соціологія культури, адже саме в цих галузях дослідники мають справу з текстами.

Перша соціологічна робота, у якій використовувалося тематичне моделювання, належала до соціології науки (Moody, Light, 2006). Однак зараз ми бачимо, що основний інтерес до цього методу присутній у соціології культури. Дослідники вважають, що соціологія культури завжди відрізнялася тим, що розвиток теорії випереджав розвиток методів.

Соціологи, які вивчають культуру, сформулювали численні теоретичні гіпотези та концепти, які обіцяють глибоке розуміння культурних змін, але їм досі не вистачає інструментів для операціоналізації концептів. Ми припускаємо, що за допомогою тематичного

моделювання буде можливо операціоналізувати такі ключові концепти, як фреймування, полісемія, гетероглосія та реляційний характер значень. (DiMaggio, Nag, Blei, 2013: 571)

Ймовірно, через те, що сенси завжди методично вивчати набагато складніше, у соціології культури був період, коли досліджувалась не стільки сама культура, скільки те, як вона виробляється (Peterson, Anand, 2004). Відсилають до смислів концепти - символічні кордони (М. Ламонт), культурні інструменти (Е. Суїдлер), когнітивні схеми (П. ДіМаджіо) та культурні фрейми (Р. Бенфорд та Д. Сноу) - розвивалися на основі «маленьких» даних, які мали на увазі техніку «повільного читання» (closereading) транскриптів інтерв'ю та проведення контент-аналізу ключових текстів (Bail, 2014).

Справді, зазвичай соціологи аналізували тексти трьома способами (DiMaggio, Nag, Blei, 2013: 577). Перший ґрунтується на інтерпретативному читанні, без будь-якої формалізації. Другий спосіб будується на контент-аналізі, при якому дослідник заздалегідь створює систему категорій та кодів, згідно з якими потім кодується текст. Обмеженням методу виявляється трудомісткість, що робить його малопридатним для аналізу великого корпусу тексту. При цьому заздалегідь потрібно добре уявляти, що можна знайти в тексті (DiMaggio, Nag, Blei, 2013: 577). І, нарешті, третя стратегія полягає в тому, щоб за допомогою програми визначити набір ключових слів, а потім порівняти, як часто у різних частинах тексту зустрічаються ці слова. Ця стратегія не зовсім влаштовувала саме соціологів, які вивчають культуру, оскільки слова витягувалися без урахування смислового контексту, у якому вони вбудовані. Дві останні стратегії більшою мірою підходять для аналізу невеликих корпусів текстів, із заздалегідь продуманими питаннями (DiMaggio, Nag, Blei, 2013: 577). Потрібен був новий метод, позбавлений колишніх недоліків. Таким методом, на думку соціологів, є тематичне моделювання, оскільки саме воно відповідає умовам аналізу великих масивів тексту.

У чому його переваги? Цей підхід має експліцитний характер, тобто масив даних доступний всім, і аналіз можна відтворити; підхід є автоматичним, що дозволяє працювати з текстами великих обсягів; він дозволяє обробляти текст до заздалегідь розробленої схеми; приймає в увагу реляційного характеру значень. В рамках тематичного моделювання корпус тексту автоматично кодується за кількома категоріями, які називають темами (topics). Алгоритм може це робити за мінімальної участі людини, тим самим метод є індуктивним за своєю природою: «Натомість щоб почати з заздалегідь визначених смислових кодів або категорій (як ті, які ми створюємо, коли кодуємо текст вручну), дослідник задає кількість тем, які повинен знайти алгоритм. Програма потім знаходить це задане кількість тем і показує ймовірність слів, що використовуються в темі, так само як надає розподіл тематик по всьому корпусу тексту» (Mohr, Bogdanov, 2013: 546).

При цьому не потрібне попереднє знайомство з текстом або заздалегідь розроблена схема кодування. Інструмент сам створює кластери, приховані теми з урахуванням статистичних моделей. Зберігається контекстуальність, так як слова приписуються кластеру з урахуванням їх появи поруч із іншими словами, як і і багатозначність смислів, оскільки слова можуть одночасно належати різним кластерам (Mutzel, 2015: 2). Незважаючи на те, що для такого дослідження потрібні знання в комп'ютерній науці та статистиці, їх неможливо проводити без людини, знайомої з тією областю, до якої належить текст. Тим самим у використанні нових інструментів для аналізу тексту спостерігається фундаментальне усунення з попередньої роботи зі створення категорій та системи кодування до інтерпретації постфактум, яка запускається, коли алгоритм знайшов тематичні категорії та потрібно вирішити, чи мають вони будь-яке значення. У цьому полягає важлива перевага такого методу, адже якщо спочатку відбувається розробка системи кодування, то коли вона закінчена і почався сам аналіз тексту, важко повернутися назад (Mohr, Bogdanov, 2013: 562). Тим самим дослідник значно більш обмежений у процедурі, і йому обов'язково потрібне глибоке знання поля ще до того, як розпочати аналіз.

Крім того, з новими інструментами дослідник може знайти тематичні категорії, про які він і не думав, що вони є в тексті, — у контент-аналізі такої можливості немає. Відповідно, є можливість досліджувати та відкривати нові патерни (DiMaggio, Nag, Blei, 2013).

**Тема 15. Програмні платформи та системи для Великих даних**

1. Системи керування потоками даних
2. Системи зберігання Великих даних
3. Платформи Великих даних
4. Обробка даних у реальному часі
5. Системи керування Великими даними
6. Аналітичні платформи

В даний час використовується значна кількість платформ та систем Великих даних. Системи обробки великих даних є фреймворками, тобто каркасами, для використання яких необхідно з'єднати їх з іншими фреймворками, прикладним програмним забезпеченням користувача та системою зберігання даних.

В аналітичному звіті BigDataAnalyticsMarketStudy, 2017 Edition наводиться така діаграма інфраструктур Великих даних, впроваджених на підприємствах, представлена у розрізі розмірів підприємств

Розподілена обробка даних тісно пов’язана з паралельною обробкою даних. Однак така обробка завжди виконується за допомогою окремих машин у кластері, підключеному до мережі. Розподілена обробка даних - це метод виконання прикладних програм групою систем. Користувач може працювати з мережевими службами та прикладними процесами, розташованими в кількох взаємопов’язаних абонентських системах. Розподілена обробка даних підвищує ефективність інформаційних потреб користувачів і забезпечує ефективність та результативність рішень.

Конкретні області у соціології культури, які можуть одержати розвиток у зв'язку з появою великих даних та нових технік аналізу, перелічені у статті К. Бейла. Серед них – картографування культурного оточення або систем значень, класифікація культурних елементів (таких як кадри або схеми всередині систем), простежування змін у культурних процесах за тривалий час. Багато питань у рамках соціології культури вимагають макроаналізу, тобто погляду згори на весь культурний простір.

Бейл закликає активно користуватися онлайновими даними в рамках соціології культури, оскільки найчастіше в руках дослідника можуть виявитися не тільки текстові дані, які цікавлять як сукупність якихось значень, але та соціальна інформація про акторов, що дозволяє ставити більш цікаві питання (Bail, 2014).

Важливо розуміти, що використання тематичного моделювання — це лише початок. Як пишуть соціологи: «В аналізі культури метою моделювання є розуміння структури даних, щоб мати можливість виявити тематичні кластери («голоси» чи «фрейми»), які ґрунтуються на даних та піддаються інтерпретації. Надалі вчені можуть використовувати їх для постановки більш фокусованих питань» (DiMaggio, Nag, Blei, 2013: 602-603). Наприклад, у процитованій роботі П. ДіМаджіо та його колег аналізувалося, як у газетних статтях представлено мистецтво. Тематичне моделювання дозволило побачити теми, потім дослідники запитали себе про зв'язок фреймування мистецтва у масовій пресі з різноманітними способами його фінансування. Для відповіді це питання вже знадобилися техніки регресійного аналізу.

Інструменти комп'ютерної науки розвивають нові методи аналізу, які зовсім не обов'язково застосовувати лише на великих даних. Вони можуть дати цікаві результати навіть на порівняно маленьких даних, які раніше аналізувалися традиційними методами. Наприклад, дослідники вважають, що комп'ютерний аналіз краще працює, ніж інтерпретативне читання. Стаття Дж. Мора та його співавторів ілюструє застосування можливостей комп'ютерного аналізу тексту до даних невеликого масштабу (Mohretal., 2013). Вони пропонують звернутися до нової стратегії комп'ютерного читання текстових повідомлень із використанням аналітичної моделі, розробленої на основі концептів Кеннета Берка. У дослідженні аналізуються тексти про стратегію національну безпеки США з 1990 по 2010 рік – це відкриті

документи, які публікуються щорічно. Автори шукали у тексті структуру риторики на глибшому рівні, ніж просте читання тесту. Тексти стратегій є не найбільшим масивом даних, дослідникубуло б під силу їх все прочитати, проте, по На думку авторів, застосування автоматичних методів дає кращі результати для виявлення риторики документа та його прагматичного контексту.

У дослідженні використовувалися три різні способи автоматичного аналізу тексту - для ідентифікації агентів та акторів застосовувався метод природного обробки мови; семантичні техніки - для пошуку «актів» через пошук присудків, пов'язаних з акторами; машинне навчання дозволило проаналізувати «сцени» у термінах Берка, у яких розташовувалися актори та його дії.

Усього було виявлено десять тематичних груп, які концептуалізувалися як «сцени» Берка (тероризм, погрози, права людини, економічний розвиток, енергія та інші). Заземлення списку знайдених акторів та їх дій дозволило сфокусовано працювати з текстом, причому не просто показати, як теми змінюються згодом, але як одні й самі актори присутні у різних тематичних групах, або дії, які спочатку виникли в одній сцені, переносяться до інших. Так, автори виявили, що після атаки 9/11 «актори» та «акти», які належали до сцени «тероризму», стали поширюватися на інші «сцени», які стосуються, наприклад, питань енергетичних ресурсів (Mohretal., 2013).

Серед основних обмежень роботи з великими даними називають доступ до них (Golder, Macy, 2014). Є ті, хто виробляє дані, — звичайні люди, які залишають електронні сліди. Є ті, хто має можливість агрегувати дані та отримати доступ до них. Але найвпливовіші — це ті, хто має змогу їх аналізувати. Обійти обмеження можна, створюючи спеціальну інфраструктуру, що, проте, потребує великих фінансових вкладень. Втім, є й простіший шлях — звернутися до аналізу даних, доступ до яких відкрито всім. Ми можемо припустити, що особливу роль нові дані будуть грати в тих областях, де немає серйозних обмежень щодо їх доступу.

Серед них соціологія культури, значна частина даних для якої можнау вигляді текстів, для аналізу яких вже є сучасні інструменти комп'ютерної науки. Безпрецедентні можливості спостереження за поведінкою людей у реальному часу залучають вчених, які не мають бекграунду у соціальних науках, але мають достатні навички для аналізу таких даних. Вони нерідко вважають, що у соціальних науках з приходом великих даних та інструментів комп'ютерної науки повинні відбутися радикальні зміни, насамперед пов'язані зі скасуванням соціальної теорії. Соціологи, ймовірно, побоюючись колонізації з боку інженерних наук, пропонують оновлений варіант соціології. На останніх сторінках огляду ми розглянемо різні варіанти майбутнього соціології як академічної дисципліни.

**Версії дисциплінарного майбутнього. Юрисдикція соціології та нові претенденти**

Позиція крайнього емпіризму представлена у статті аналітика Кріса Андерсона, який раніше очолював журнал "Wired". У 2008 році він проголосив «кінець теорії» та необхідність відмовитися від наукового методу в його колишньому вигляді.

Зараз існує найкращий шлях. Петабайти дозволяють нам сказати: «Досить з нас кореляцій». Ми можемо аналізувати дані без гіпотез про те, які зв'язки мають бути присутніми. Ми можемо помістити всі ці цифри у найбільші комп'ютери, які тільки відомі світу, і дозволити статистичного алгоритму знайти патерн там, де його не бачить наука. реляція зайняла місце каузальності, і наука може розвиватися навіть без когерентних моделей, уніфікованих теорій чи будь-якого існуючого механічного пояснення. Немає жодної причини чіплятися за минуле. (Anderson, 2008, цит. по: Kitchin, 2014: 4)

З цього погляду соціальні науки має змінити нову атеоретичну науку про дані. Соціологи та політологи повинні поступитися своїм місцем аналітикам даних, які обтяжені теоретичним багажем соціальних наук. Аналітики часто переконані, що можна обійтися без заздалегідь продуманих теорій, моделей чи гіпотез — алгоритми можуть змусити «дані говорити самі за себе». Якщо раніше дослідники були потрібні для генерації даних, то зараз в цьому немає необхідності. Дисциплінарна компетенція менш важлива проти технічними навичками. Дані зможуть дати відповіді на будь-які питання лише після певних комп'ютерних маніпуляцій, відповідно, необхідною є вчений ступінь у галузі комп'ютерних наук, а не в соціології.

Про загрозу юрисдикції соціології писали ще до того, як великі дані підірвали Інтернет. У 2007 році вийшла стаття з назвою, яка говорить саме за себе: «Наступаючий криза емпіричної соціології» (Burrows, Savage, 2007). Її автори були стурбовані тим, що комерційні компанії мають справу з даними, які мріють отримати багато соціологів. Комерційна соціологія, на їхню думку, існує як відповідь на рефлексивний характер сучасного капіталізму, якому потрібні знання та інформація, щоб отримувати ще більше прибутку. Вже тоді соціологи побачили в цьому небезпеку: «Ми були стурбовані, оскільки розцінили це як ще один цвях у кришку труни академічної соціології та її домагань на юрисдикцію знання про соціальне» (Burrows, Savage, 2014: 2). Раніше методи робили свій внесок в унікальний характер дисципліни, зараз дані можуть з'явитися без розрахунку вибірки, проведення інтерв'ю чи фокус-групи.

Інформація про визнаних дослідників у галузі аналізу великих даних дає можливість побачити, наскільки серйозні побоювання щодо колонізації соціальних дисциплін інженерними науками Список було отримано на основі програм кількох ключових конференцій, що проводилися в області обчислювальної соціальної науки. Цей список не

претендує на те, щоб бути вичерпним у цій галузі, проте його достатньо, щоб оцінити кількість дослідників крім соціологів. Тут потрібно звернути увагу на галузь знання, в якій учасниками було отримано науковий ступінь: трохи менше половини - в галузі соціальних наук, інша половина захистила дисертації у галузі природничих, інженерних та комп'ютерних наук. Зараз вони афілійовані не тільки з різними університетськими структурами, але і з індустрією (Facebook, Microsoft). Власне університетські департаменти також представлені приблизно в рівних пропорціях: в галузі соціальних наук їх трохи менше - 17, з них 7 - з соціології. За невеликим винятком, якщо дослідник має вчений ступінь у галузі технічних чи природничих наук, то і працюватиме надалі він також у структурах у межах цих напрямів.

Поділ різні галузі знання зберігається й у публікаціях. Звісно, у виборі журналу автори вільніші, ніж у виборі місця роботи. Автори нашої вибірки часто публікували статті в міждисциплінарних журналах. При цьому у соціологічних виданнях переважно з'являються дослідники із соціальних наук. Список основних журналів: "SocialNetworks" - 17, "AmericanJournalofSociology" - 7, "SocialForces" - 4, "AmericanSociologicalReview", "JournalofMathematicalSociology», «SocialScienceResearch» – по 3 статті. Ми не бачимо ні одного автора з нашого списку, який здобув би технічну освіту, на сьогодні працює у профільному департаменті і при цьому публікується у соціологічних журналах. Усі автори без соціологічного бекграунду, які вивчають соціальну поведінку, обирають для публікацій престижні міждисциплінарні видання - "Nature", "Science", "PlosOne", "PNAS", "ScientificReports".

Отже, можна говорити, що юрисдикція соціології справді оспорюється із боку інших галузей знання. Однак поки що це не зачіпає власне соціологічні робочі позиції та журнали, концентруючись у спеціальному просторі, призначеному для міждисциплінарних досліджень. Через це більшість соціологів може помічати процес колонізації чи надавати йому серйозного значення. Але є кілька винятків, про які далі йтиметься. Доказова соціальна наука та її заклик «йти від даних» Дослідники вважають, що виробництво знання в соціології має змінитися через те, що інші дисципліни також почали використовувати дані про соціальні транзакції. Як пише Кінг, «зараз соціальні науки зазнають історично важливі зміни, коли їх більшість рухається від виробництва знання, властивого гуманітаристиці, до природничих наук у тому, що стосується дослідницького стилю, інфраструктури, доступності даних, емпіричних методів, змістовного розуміння та можливості для швидкого та помітного зростання» (King, 2013: 165). Про які зміни йдеться? Панівний нині науковий стиль американської соціології сформувався до кінця 1970-х років, і зараз він домінує на сторінках провідних соціологічних журналів.

Головна його відмінність - це застосування опитувальних інструментів та статистики для перевірки заздалегідь сформульованих гіпотез Вважається, що у дослідженнях американської соціології теорія передує етапу збору даних, який спрямований на пошук статистичної підтримки заздалегідь сформульованих гіпотез (McFarland, Lewis, Goldberg, 2015; Pontille, 2003). Початкове формулювання гіпотез при такому стилі має велике значення, адже немає можливості зібрати будь-які дані. Оскільки зараз дані з'являються не внаслідок зусиль дослідників, відповідно, очікується змін у прийнятому порядку процесів соціологічного дослідження.

Дедалі більше можна зустріти робіт, присвячених протиставленню теоретико-орієнтованої науки (theory-driven) дослідженням, які займають емпірицистську позицію "йти від даних" (data-drivenscience). У таких дослідженнях гіпотези можуть виникнути з доступних даних (Kitchin, 2014: 6). Ці роботи закликають перестати вдавати, що гіпотези формулюються до того, як дослідження було розпочато та закінчено (Goldberg, 2015; McAbee, Landis, Burke, 2017). У багатьох випадках гіпотези з'являються у міру проведення дослідження, але у підсумковому тексті дослідник створює ілюзію, що гіпотези з'являються на основі всіх прочитаних джерел та спрямовують дії дослідника. А. Голдберг справедливо пише про те, що навряд чи знайдеться дослідження, автор якого зізнається, що поки він шукав відповіді на одне запитання, знайшов відповідь на зовсім іншу. Існуючий формат статті задає логіку лінійного викладу, який випливає з прийнятих норм. Дані повинні отримати необхідне теоретичне оформлення, що доводить, що керували гіпотези перебігом дослідження.

Це досить витончено продемонстрував М. Теплицький, коли порівняв, як змінювалися тексти соціологів від варіанта розгорнутої доповіді на конференції до статті у науковому журналі (Teplitskiy, 2016). Крім іншого, він мав можливість побачити, на що найчастіше спрямована критика рецензентів, змінюють чи вони теорію чи його зауваження більшою мірою ставляться до аналізу даних.

**Тема 16. Машинне навчання за допомогою бібліотеки Scikit-learn.**

1. Види машинного навчання.
2. Основні бібліотеки машинного навчання Python (Scikit-learn, Keras, TensorFlow). Вибір найкращої моделі.
3. Створення моделі. Вивчення моделі. Тестування моделі.
4. Функціонал бібліотеки Scikit-Learn.

Види машинного навчання. Основні бібліотеки машинного навчання Python (Scikit-learn, Keras, TensorFlow). Створення тренувальних наборів - передобробка даних. Точність та достовірність моделі. Вибір найкращої моделі.

Кроки типового практичного сценарію машинного навчання. Завантаження набору даних. Дослідження даних за допомогою Pandas. Візуалізація ознак за допомогою Matplotlib. Розбиття даних для навчання та тестування. Створення моделі. Вивчення моделі. Тестування моделі.

Налаштування параметрів моделі та оцінка її точності. Формування прогнозів на на підставі «живих» даних, які ще невідомі моделі.

Функціонал бібліотеки Scikit-Learn. Класифікація за допомогою К-сусідів.

Лінійні моделі для регресії та класифікації (модель лінійної регресії, логістична регресія, та ін). Наївні байєсівські класифікатори. Дерева рішень та випадковий ліс. Спосіб опорних векторів. Основи нейронних мереж.

Метод основних компонентів. Алгоритми кластеризації (кластеризація методом К-середніх, ієрархічна кластеризація, та ін).

Якби соціологічні дослідження справді запускалися теоретично обґрунтованим питанням, Теплицький навряд чи виявив би, що після процедури рецензування головним чином змінюється теоретичне обрамлення статті, тоді як аналіз даних залишається без помітних змін. Здавалося б, між теорією, дослідницьким питанням, даними та аналізом має існувати більш-менш стійка зв'язок, відповідно, у разі зміни теорії має змінитися аналіз даних. Однак у більшості соціологічних статей теорія змінюється, а аналіз залишається тим самим, що дозволяє говорити швидше про теоретичне обрамлення, ніж про повноцінну опору на теорію.

Можливо, роль даних та їх аналізу і раніше була самостійнішою у дослідженні, чим це фіксувалося на риторичному рівні. Головним фактором, чому дослідження відбулося, цілком міг бути доступ до даних, а не зазор у теоретичному знанні, який і підказав ідею дослідження. Але саме зараз вчені закликають відмовитись від Sharking (SecretlyHypothesizingAfterResultsAreKnown) і почати слідувати Tharking (TransparentlyHypothesizingAfterResultsAreKnown), тобто припинити приховувати, як здійснювалося дослідження (Hollenbeck, Wright, 2016). Важливо, що Tharking не є data-mining, колибез усяких ідей вивчаєш дані та просто отримуєш закономірності. Мова йде про появу гіпотез, коли дані виявляють нові патерни, дають нові ідеї для міркувань.

У тому, щоб емпірико-орієнтована соціальна наука стала легітимнішою, може сприяти використання нових даних. В силу того, що дані створюються без дослідника, у яких виявляється великий потенціал саме для індуктивного способу аналізу (McAbee, Landis, Burke,

2017). Автори пишуть, що немає необхідності протиставляти такі дослідження дедуктивному способу, швидше потрібно прагнути їхньої більшої легітимності. Для їх позначення використовується спеціальна назва - доказова соціальна наука (forensicsocialscience), у якій мають поєднатися дедуктивний та індуктивний підходи. Дослідники не повинні займатися перевіркою даних на наявність всіх можливих зв'язків, вони також не повинні повністю фокусуватися на перевірці гіпотез, тому що можна упустити несподівані емпіричні знахідки. Для того щоб доказова соціальна наука стала повноцінною наукою, яка створює та розвиває теорії, «дослідники повинні працювати з даними, знаходити важливі патерни, а потім робити крок назад до побудови осмислених аналітичних конструктів» (McFarland, Lewis, Goldberg, 2015).

Соціологів не вражають лише патерни, тому вони готові внести корективи до дослідницького стилю соціології, проте не збираються відмовлятися від необхідності розвивати соціальну теорію та пропонувати соціологічні пояснення. Нові дані можуть бути корисними для виявлення закономірностей, але головне в соціальних науках — це їхнє пояснення. У такому у випадку великі дані можуть дати ті самі емпіричні загадки, які повинні бути присутніми в дослідженнях: «Методи великих даних не є кінцевою метою, вони лише частина руху до пояснювальної теорії» (Halavais, 2015: 587).

Дослідник може не знати заздалегідь, який він виявить патерн на основі великих даних, проте надзвичайно важливо, щоб його дослідницькі амбіції диктували йому не зупинятися лише на одному патерні. Наприклад, техніки тематичного моделювання використовувалися для реконструкції зв'язків між дисциплінами через аналіз імпорту та експорту мови один одного. Зв'язки будувалися на основі 1 000 000 дисертацій, написаних з 1980 по 2010 рік у 157 американських університетів. Автори виявили, що методологічні (статистика, математика), технологічні (комп'ютерні науки) та абстрактні тематичні категорії працюють на експорт – їх досягнення використовуються у ряді інших дисциплін, тоді як самі ці області замкнуті та рідко запозичують мову інших наук. Було також підраховано кількість слів, що стосуються внутрішньої та зовнішньої мови. Виявилось, що соціологія згодом демонструє помітне зниження частки внутрішньої мови та збільшення зовнішньої. На основі цього автори зробили висновок у тому, що соціологія є типом науки, яка завжди залишається на стадії відкриттів. Ця стадія характеризується більш помітною роллю зовнішньої мови. Інші науки також можуть спиратися на зовнішню мову, проте їхній власний продовжує активно розвиватися (Macfarlandetal., 2013).

Дослідження виконано з урахуванням великих текстових даних, які аналізуються просунутими інструментами. Це чудова можливість отримати цікаві результати про зміни мови дисциплін та їх зв'язків один з одним.

Результати можуть стати відправною точкою, приводом поставити питання, чому домінує зовнішню чи внутрішню мову або чому їх співвідношення змінюється з часом. Як пише Китчин: «Одна річ — знайти патерн, інша — її пояснити. Це потребує глибокого знання соціальної теорії та контексту. Фактично, патерн — це кінцева точка, а початкова для додаткового аналізу, який майже напевно вимагатиме нових даних» (Kitchin, 2014: 8).

Існує і радикальніша пропозиція — розгорнути соціологію від каузальних пояснень у бік описів. М. Севідж та Р. Берроуз не просто закликають інкорпорувати великі дані у свої роботи, але пропонують зайнятися дослідженнями, в яких буде більше патернів, ніж пояснень (Savage, 2009; Burrows, Savage, 2007). Соціологи повинні серйозно замислитись про причинизгасаючого інтересу широкої публіки до власних досліджень. Пропозиція Севіджа та Берроуза полягає у відмові від каузальності, оскільки соціології так і не вдалося запропонувати переконливих пояснень. Краще, ніж соціологія може зараз зайнятися, це робити хороші описи, використовуючи нові методи та дані. У своїх міркуваннях Севідж спирається на ідеї Е. Ебботта про дисциплінарний проект, в рамках якого соціологія могла б існувати без того, щоб ставити каузальність на перше місце, як це можливо в інших дисципліни:

Одна з головних причин, чому публіка перестала цікавитись соціологією, це наше поблажливе ставлення до опису. Публіка жадає описи, але ми надто зневажаємо цей жанр. Зосереджуючись на одній тільки каузальності, ми відмовляємо у публікації статтям з чистим описом, навіть якщо опис виконано з використанням кількісних методів та має важливі змістовні висновки. У той же час, комерційні фірми платять мільйони за таку роботу, виходить, що наше суспільство фактично «описується» найдокладнішим чином приватними маркетинговими компаніями. Але ми, хоч і любимо вважати, що відповідальні за публічне знання про суспільство, зневажаємо і описи та методи, які зазвичай використовуються для кількісних досліджень. Наші соціальні індикатори є майже випадковим набором змінних, придатних для каузального аналізу. (Abbott, 2001: 121) Для Ебботта соціологія ніколи не сприйматиметься всерйоз як наука про соціальне життя, доки вона не візьметься за описи. Соціологія все ще не зробила повний розворот у цей бік, проте, можливо, великі дані наблизять його. Перший крок у цей бік було зроблено, коли вчені почали міркувати про дослідження, що йде від даних із нелінійною послідовністю кроків, як про легітимний варіант дисциплінарного майбутнього соціології.

Висновок

Не складно знайти безліч прикладів корисності великих даних для зовнішнього світу. Однак чи можна вже говорити про досягнуті успіхи в соціології? Є думка, що поки що ми частіше маємо справу з міркуваннями про застосування великих даних у соціальних науках,

ніж справді з дослідженнями, що будуються на їхньому аналізі (Halavais, 2015). Соціологи досі частіше критикують можливості більших даних, ніж їх використовують. Серед десяти найуживаніших. Знайдені передбачення можуть дозволити досягти поліпшення у комерційній сфері. Один із найбільш цитованих прикладів — використання пошукових запитів для інформації про реальне поширення захворювань. Зазвичай у Європі та США інформація про грип збирається на основі візитів до лікаря, дані публікуються щотижня із запізненням у 1–2 тижні. Пошукові запити дають можливість відслідковувати захворювання швидше, причому можна заздалегідь отримати інформацію, яка відповідатиме реальній поведінці (Ginsbergetal., 2009).

Ключових слів у статтях, присвячених великим даним, половина відноситься для обговорення викликів та можливостей їх використання — challenge, revolution, opportunity, value, application, future9. Можливо, варто погодитися, що змістовний прорив та важливі наукові результати чекають на нас попереду. З іншого сторони, саме зараз важливо обговорити, що може змінитися у соціології при переорієнтації дослідника зі збору даних на постановку питань до існуючих масивів.

Найбільшу увагу дослідників поки що зосереджено на тому, щоб визначити відмінності нової соціальної науки, дослідження в якій далеко не завжди носять лінійний характер. Набагато рідше обговорюється питання про епістемологічний статус нового типу даних. У статтях майже за умовчанням вважається, що великі дані - справжні та об'єктивні. Якщо вони не створювалися за запитом дослідника, то нібито мають більшу надійність. Однак варто пам'ятати, що у разі великих даних «дослідник не лише позбавлений можливості впливати на інструмент, але й нерідко не може спостерігати його у дії». У створенні нових даних велику роль відіграють електронні механізми, які, зауважимо, створюються та обслуговуються людьми.

У цьому сенсі як ніколи важливо продовжити ставити питання, з якими ж ми маємо справу і про що вони можуть нам розповісти. Таким чином, один з необхідних кроків дослідження має полягати у критичній оцінці виробництва даних, що допоможе уникнути ситуації виявлення та опису хибних залежностей.

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ «ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Кафедра соціології і публічного управління
(назва кафедри, яка забезпечує викладання дисципліни)

# ЗАВДАННЯ ДЛЯ САМОСТІЙНОЇ РОБОТИ З НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

## МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ
( назва навчальної дисципліни)

рівень вищої освіти другий (магістерський)
перший (бакалаврський) / другий (магістерський)

галузь знань 05 Соціальні та поведінкові науки
(шифр і назва)

спеціальність 054 Соціологія
(шифр і назва )

освітня програма Соціологічне забезпечення економічної діяльності
(назви освітніх програм спеціальностей )

вид дисципліни професійна підготовка; обов'язкова
(загальна підготовка / професійна підготовка; обов'язкова/вибіркова)

форма навчання денна
(денна / заочна/дистанційна)

Харків – 2024 рік

## Тема 1. Основні елементи формалізму

1. Проблеми неодновимірності багатьох досліджуваних соціологом понять.
2. Особливості вивчення простору сприйняття соціологічних явищ та процесів – основне завдання БШ.
3. Ідеї Кумбса щодо урахування можливості упорядкування відстаней між об'єктами.
4. Векторна модель або модель ідеальної крапки як основа БШ.
5. Функція відстані (аксіоматичне визначення).
6. Відповідні функції стресу.

**Література:** 1, 3, 4, 5

## Тема 2. Багатовимірне розгортання та індивідуальне багатовимірне шкалірування

1. Постановка завдання важливість врахування специфіки метрик окремих респондентів.
2. Вид вхідних і вихідних даних, функції стресу в індивідуальному БШ.
3. Одномірне розгортання.
4. Обґрунтування необхідності переходу до простору довільної розмірності для успішного виконання завдання шкалірування.
5. Неметричне багатовимірне розгортання.
6. Особливості інтерпретації результатів.
7. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5, 7

## Тема 3. Проблеми формування вихідних даних і інтерпретації результатів у багатовимірному шкалірування

1. Роль соціолога при отриманні даних, вихідних для багатовимірного шкалірування та інтерпретації його результатів.
2. Класифікація відповідних способів опитування; проблеми, що постають при такому способі збору даних.
3. Приклади розрахунку матриці близькості на основі аналізу достатньо надійних даних іншого роду.
4. Використання формальних та неформальних методів при інтерпретації результатів багатовимірного шкалірування. Значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей.
5. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5, 6

**Тема 4. Канонічний аналіз. Загальне уявлення про методи, які засновані на моделях частот**

1. Загальне уявлення про моделювання частот таблиці спряженості.
2. Мультиплікативні та адитивні моделі частот.
3. Роль логарифмування мультиплікативної моделі.
4. Основне завдання канонічного аналізу. Принципи їх отримання на основі аналізу таблиці спряженості.
5. Моделі частот, що відповідають канонічному аналізу.
6. Зв'язок канонічних коефіцієнтів кореляції з критерієм «хі-квадрат".
7. Загальне уявлення про оцифрування значень номінальних ознак.
8. Канонічний аналіз як метод оцифровки і метод вимірювання зв'язку між двома номінальними ознаками зі "спільними альтернативами".
9. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5, 6

**Тема 5. Логлінейний аналіз**

1. Причини відмінності реального розподілу від рівномірного.
2. Моделі частот, що відповідають логлінейному аналізу.
3. Насичена модель.
4. Мета переходу до логарифмів частот.
5. Гіпотези про взаємозв'язок ознак. Їх роль при побудові моделей частот.
6. Розрахунок коефіцієнтів логлінейной моделі для двовимірного випадку. Відносини переважання. Інтерпретація коефіцієнтів через відносини переважання (для моделі довільної розмірності).
7. Порівняння логлінейного аналізу з номінальним регресійним і дисперсійним аналізом, а також з методом послідовних розбивок. Порівняння здійснюється на змістовному рівні.
8. Різне розуміння залежної ознаки: кількісна ознака в дисперсійному аналізі, кількісна або номінальна – в номінальному регресійному і частота, що стоїть в клітці багатовимірної таблиці спряженості, – в логлінейном аналізі.
9. Розв’язання практичних завдань.

**Література:** 1, 2, 3, 4, 5

**Тема 6. Причинний аналіз. Стратегія аналізу структури взаємозв'язків ознак**

1. Поняття причини в соціології. Принципова неможливість повністю його формалізувати.
2. Граф причинних зв'язків.

3. Повторення принципів побудови часткових коефіцієнтів кореляції і регресії. Важливість для соціолога вивчення відповідних зв'язків.
4. Поняття "помилкової" кореляції. Основні причинні схеми, що призводять до їх появи.
5. Координуючий шлях. Його ефект.
6. Обчислення ковариаций (кореляцій) між будь-якими двома ознаками на основі графа зв'язків.
7. Структурні рівняння.
8. Обчислення структурних коефіцієнтів. Їх зв'язок з частковими коефіцієнтами регресії.
9. Основна теорема причинного аналізу. Її роль у вивченні статистичних залежностей.
10. Поняття структури багатовимірної випадкової величини.
11. Формування узагальнених показників на базі аналізу структури зв'язків ознак.
12. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5, 6

## Тема 7. Завдання розпізнавання образів. Поняття автоматичної класифікації об'єктів

1. Класифікація як один із фундаментальних процесів у науці.
2. Загальне уявлення про завдання розпізнавання образів (синоніми: образ, клас, кластер, таксон; неоднозначність трактування термінів в літературі).
3. Виділення завдань: пошук класів, опис класів, визначення найбільш ефективної системи ознак.
4. Виділення задачі автоматичної класифікації об'єктів (синоніми: багатовимірна класифікація, розпізнавання образів без вчителя, кластерний аналіз, таксономія).
5. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5, 6

## Тема 8. Проблема "стикування" змісту і формалізму при використанні алгоритмів класифікації

1. Специфіка рішення соціологічних завдань побудови типології за допомогою методів автоматичної класифікації.
2. Сенс протиставлення термінів "класифікація" і "типологія".
3. Виділення основних формальних елементів алгоритмів автоматичної класифікації, що вимагають стикування зі змістовними концепціями соціолога.

4. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5, 6

## Тема 9. Функції відстані між об'єктами

1. Обумовленість гіпотез апріорними уявленнями дослідника про типи об'єктів.
2. Загальне уявлення про розмиті класифікації.
3. Доцільність комплексного використання декількох алгоритмів класифікації в соціологічних завданнях побудови типології.

**Література:** 1, 2, 3, 4, 5

## Тема 10. Основні види процедур класифікації. Відстані між класами

1. Причини необхідності розгляду відстаней між класами в ієрархічних процедурах.
2. Алгоритм найближчого сусіда як приклад способу класифікації, що використовує такі відстані.
3. Приклади соціологічних задач, для яких змістовно адекватні різні способи вимірювання відстаней між класами.

Література: 1, 3, 4, 5, 7

## Тема 11. Гіпотези про розташування об'єктів у ознаковому просторі

1. Обумовленість гіпотез апріорними уявленнями дослідника про типи об'єктів.
2. Основні види гіпотез: компактності, зв'язності (безперервності), унімодального розподілу.
3. Приклади алгоритмів, що шукають закономірності розташування точок у ознаковому просторі, що відповідають кожній з гіпотез: алгоритм Форель (гіпотеза компактності), алгоритм найближчого сусіда (гіпотеза зв'язності), алгоритм, заснований на виділенні локальних максимумів функції приналежності (гіпотеза унімодального розподілу ).
4. Доцільність комплексного використання декількох алгоритмів класифікації в соціологічних завданнях побудови типології.
5. Коригування результатів класифікації з метою забезпечення відповідності класифікації і типології.

**Література:** 1, 3, 4

## Тема 12. Поняття інтерпретації вихідних даних і основні методологічні принципи використання методів аналізу даних в соціології

1. Основні фактори, що визначають інтерпретацію вихідних даних: апріорні уявлення дослідника про спосіб породження цих даних (у тому числі – про моделі сприйняття респондентами пропонованих ним питань, об'єктів, про ймовірнісну природу даних і т. д.); мета дослідження; концептуальні уявлення соціолога про досліджуване явище; характер моделі явища, "закладеної" в математичному методі, використання якого планується; розгляд спостережуваних змінних як непрямих показників латентних факторів, насправді цікавлять дослідника і т. п.

**Література:** 1, 3, 4, 5

## Тема 13.Дані. Метадані

Підготувати відповіді на питання:
1. Що таке дані?
2. Які ДСТУ з визначеннями даних вам відомі?
3. Які визначення даються у ФЗ-149?
4. Що таке життєвий цикл даних?
5. Перерахуйте етапи життєвого циклу даних.
6. Для яких цілей потрібен етап “Синтез даних” (один із етапів життєвого циклу даних)?
7. З якою метою потрібен етап “Використання даних” (один із етапів життєвого циклу даних)?
8. З якою метою потрібен етап “Публікація даних” (один із етапів життєвого циклу даних)?
9. З якою метою потрібен етап “Архівація даних” (один із етапів життєвого циклу даних)?
**Література:** 4, 8, 9

## Тема 14. Великі дані. Системи керування великими даними

Підготувати відповіді на питання:

1. Що таке Великі дані?
2. Які п'ять характеристик притаманні Великим даним?
3. Які є базові принципи обробки Великих даних?

4. Оцініть скільки необхідно для зберігання набору даних, що містить координати, швидкості та метаінформацію (тип молекули та час вимірювання по конкретній молекулі) для всіх молекул біля аеропорту.
5. Що таке стовпцеві бази даних?
6. Що таке сховища документів?
7. Що таке потокові дані?
8. Що таке сховища для ключів?
9. Що таке SQL на Hadoop?
10. Що таке новий SQL?
11. Що таке графові бази даних?
**Література:** 4, 8, 9

## Тема 15. Програмні платформи та системи для Великих даних

Підготувати відповіді на питання:
1. Які мови програмування використовуються для роботи з фреймворками даних?
2. Чи дозволяє ліцензія Apache 2.0, під якою випущено деякі фреймворки, вносити власні виправлення в програмний код забезпечення?
3. Перерахуйте кілька фреймворків, які забезпечують обробку даних у реальному часі.
4. Перерахуйте кілька фреймворків, які забезпечують аналітичну обробку даних.
5. Перерахуйте кілька фреймворків, які забезпечують зберігання даних.
6. Перерахуйте кілька фреймворків, які забезпечують керування потоками даних.
**Література:** 4, 8, 9

## Тема 16. Машинне навчання за допомогою бібліотеки Scikit-learn.

Підготувати відповіді на питання:
1. Яке обладнання потрібне для обробки великих даних?
2. Центр обробки даних якого рівня забезпечує максимальну надійність?
3. Центр обробки даних якого рівня забезпечує резервування?
4. Центр обробки даних якого рівня дозволяє проводити обслуговування обладнання одночасно з обробкою даних?
5. Як багато часу потрібно для створення центру обробки даних "під ключ"?
**Література:** 4, 8, 9

Основна література
1. Горбачик А.П., Сальнікова С.А. Аналіз даних соціологічних дослідженьзасобами SPSS: Навч. посіб.- Луцьк, 2008. – 164 с. IBM SPSS 20 інструкція користувача// https://www.xn--80aaexjatkpdgggghih8b1a2yhv.com.ua/ibm/spss-20/%D1%96%D0%BD%D1%81%D1%82%D1%80%D1%83%D0%BA%D1%86%D

1%96%D1%8F-%D0%BA%D0%BE%D1%80%D0%B8%D1%81%D1%82%D1%83%D0%B2%D0%B0%D1%87%D0%B0.
2. Паніотто В.І., Максименко В. С., Харченко Н.М. Статистичний аналізсоціологічних даних. - Київ, 2004. – 270 с. Литвин В.В. Аналіз даних та знань: підручник/ В.В. Литвин, В.В. Пасічник, Ю.В. Нікольський.- Л.: Магнолія, 2020.- 276с. (базовий підручник).
Допоміжна література
3. Лупан І.В., Авраменко О.В., АкбашК.С.Комп’ютерні статистичні пакети: навчально-методичнийпосібник. - 2-е вид. - Кіровоград: ’КОД”. 2015. - 230 с. - http://dspace.cuspu.edu.ua/jspui/bitstream/123456789.
MakingSenseofMultivariateData Analysis//https://us.sagepub.com/en-us/nam/book/making-sense-multivariate-data-analysis
4. Бахрушин В.Є.Методи аналізу даних: навчальний посібник для студентів В.Є. Бахрушин. - Запоріжжя : КПУ, 2011. - 26В с. - http://web.kpi.kharkov.ua/auts/wp-content/uploads/sites/67/2017/02/DAMAP_Ivashko_posobie2.pdf
5. Інтелектуальний аналіз даних: практикум/ М.Т. Фісун, І.О. Кравець, П.П. Казмірчук.- Л.: Новий Світ-2000, 2020.- 162с. Гладун А.Я., Рогушина Ю. В. DataMining: пошук знань в даних. Київ. ТОВ «ВД «АДЕФ- Україна», 2016. — 452 с..

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ**
**«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Кафедра соціології і публічного управління
(назва кафедри, яка забезпечує викладання дисципліни)

# ПЛАНИ СЕМІНАРСЬКИХ ЗАНЯТЬ З НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

# МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ

( назва навчальної дисципліни)

рівень вищої освіти другий (магістерський)
перший (бакалаврський) / другий (магістерський)

галузь знань 05 Соціальні та поведінкові науки
(шифр і назва)

спеціальність 054 Соціологія
(шифр і назва )

освітня програма Соціологічне забезпечення економічної діяльності
(назви освітніх програм спеціальностей )

вид дисципліни професійна підготовка; обов'язкова
(загальна підготовка / професійна підготовка; обов'язкова/вибіркова)

форма навчання денна
(денна / заочна/дистанційна)

Харків – 2024 рік

## <u>Тема 1.</u> Основні елементи формалізму (2 год.)

1. Проблеми неодновимірності багатьох досліджуваних соціологом понять.
2. Особливості вивчення простору сприйняття соціологічних явищ та процесів – основне завдання БШ.
3. Ідеї Кумбса щодо урахування можливості упорядкування відстаней між об'єктами.
4. Векторна модель або модель ідеальної крапки як основа БШ.
5. Функція відстані (аксіоматичне визначення).
6. Відповідні функції стресу.
7. Простір сприйняття респондентами запропонованих їм об'єктів.
8. Формальне визначення близькості.
9. Вихідні дані для БШ – матриця близькості між об'єктами.
10. Метричне та неметричне БШ.
11. Формальні аспекти проблем розмірності шуканого евклідового простору і обертання, що визначають його осей координат.
12. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5

## Тема 2. Багатовимірне розгортання та індивідуальне багатовимірне шкалірування (2 год.)

1. Постановка завдання важливість врахування специфіки метрик окремих респондентів.
2. Вид вхідних і вихідних даних, функції стресу в індивідуальному БШ.
3. Одномірне розгортання.
4. Обґрунтування необхідності переходу до простору довільної розмірності для успішного виконання завдання шкалірування.
5. Неметричне багатовимірне розгортання.
6. Особливості інтерпретації результатів.
7. Спосіб обліку таких метрик в індивідуальному БШ.
8. Модель ідеальної точки в багатовимірному випадку.
9. Функція стресу.
10. Специфіка вихідних даних (наявність двох видів точок, що відповідають об'єктам і респондентам відповідно).
11. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5, 7

## Тема 3. Проблеми формування вихідних даних і інтерпретації результатів у багатовимірному шкалірування (2 год.)

1. Роль соціолога при отриманні даних, вихідних для багатовимірного шкалірування та інтерпретації його результатів.
2. Класифікація відповідних способів опитування; проблеми, що постають при такому способі збору даних.
3. Приклади розрахунку матриці близькості на основі аналізу достатньо надійних даних іншого роду.
4. Використання формальних та неформальних методів при інтерпретації результатів багатовимірного шкалірування. Значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей
5. Можливі способи одержання вихідних даних.
6. Проблеми застосування статистичних методів в соціології.
7. Основні функції та процедури аналізу даних.
8. Значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей.
9. Створення багатовимірних таблиць за допомогою вторинних змінних.
10. Загальна характеристика сучасних програмних засобів аналізу соціологічних даних.
11. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5, 6

## Тема 4. Канонічний аналіз. Загальне уявлення про методи, які засновані на моделях частот (2 год.)

1. Загальне уявлення про моделювання частот таблиці спряженості.
2. Мультиплікативні та адитивні моделі частот.
3. Роль логарифмування мультиплікативної моделі.
4. Основне завдання канонічного аналізу. Принципи їх отримання на основі аналізу таблиці спряженості.
5. Моделі частот, що відповідають канонічному аналізу.
6. Зв'язок канонічних коефіцієнтів кореляції з критерієм «хі-квадрат".
7. Загальне уявлення про оцифрування значень номінальних ознак.
8. Канонічний аналіз як метод оцифровки і метод вимірювання зв'язку між двома номінальними ознаками зі "спільними альтернативами".
9. Поняття зв'язку між двома групами ознак.
10. Послідовність канонічних коефіцієнтів кореляції.
11. Принципи отримання канонічних коефіцієнтів кореляції на основі аналізу таблиці спряженості.
12. Використання канонічної кореляції в аналізі таблиць спряженості.
13. Необхідність сполучення моделі, закладеної в конкретному методі оцифровки.
14. Побудова соціологічних індексів за допомогою техніки канонічного аналізу.

15. Вирішення проблеми зважування складових індекс ознак.
16. Розв’язання практичних завдань.
**Література:** 1, 2, 3, 4

## Тема 5. Логлінейний аналіз (2 год.)

1. Причини відмінності реального розподілу від рівномірного.
2. Моделі частот, що відповідають логлінейному аналізу.
3. Насичена модель.
4. Мета переходу до логарифмів частот.
5. Гіпотези про взаємозв'язок ознак. Їх роль при побудові моделей частот.
6. Розрахунок коефіцієнтів логлінейной моделі для двовимірного випадку. Відносини переважання. Інтерпретація коефіцієнтів через відносини переважання (для моделі довільної розмірності).
7. Порівняння логлінейного аналізу з номінальним регресійним і дисперсійним аналізом, а також з методом послідовних розбивок. Порівняння здійснюється на змістовному рівні.
8. Різне розуміння залежної ознаки: кількісна ознака в дисперсійному аналізі, кількісна або номінальна – в номінальному регресійному і частота, що стоїть в клітці багатовимірної таблиці спряженості, – в логлінейном аналізі.
9. Неможливість отримання нового знання на основі аналізу рівномірного розподілу (суть аналізу даних – вивчення змін, порівняння показників різного роду).
10. Сенс вкладів різної розмірності.
11. Роль критерію "хі-квадрат" при використанні логлінейного аналізу.
12. Відносини переважання. Інтерпретація коефіцієнтів через відносини переважання (для моделі довільної розмірності).
13. Різні можливості пошуку поєднань значень предикторів: перевірка гіпотез про наявність багатовимірних зв'язків у логлінейном аналізі і можливість пошуку найбільш дієвих поєднань в методі послідовних розбивок і регресійному аналізі, заздалегідь заданий набір поєднань значень предикторів в дисперсійному аналізі.
14. Розв’язання практичних завдань.
**Література:** 1, 2, 3, 4, 5

## Тема 6. Причинний аналіз. Стратегія аналізу структури взаємозв'язків ознак (2 год.)

1. Граф причинних зв'язків.

2. Повторення принципів побудови часткових коефіцієнтів кореляції і регресії. Важливість для соціолога вивчення відповідних зв'язків.
3. Поняття "помилкової" кореляції. Основні причинні схеми, що призводять до їх появи.
4. Обчислення ковариаций (кореляцій) між будь-якими двома ознаками на основі графа зв'язків.
5. Структурні рівняння. Обчислення структурних коефіцієнтів. Їх зв'язок з частковими коефіцієнтами регресії.
6. Основна теорема причинного аналізу. Її роль у вивченні статистичних залежностей.
7. Поняття структури багатовимірної випадкової величини.
8. Формування узагальнених показників на базі аналізу структури зв'язків ознак.
9. Роль статистичних методів при вивченні причинних відносин.
10. Структурні коефіцієнти. Вхідні (зовнішні, незалежні) і вихідні (внутрішні, залежні) змінні.
11. Правила редукції причинних схем та формування рівнянь.
12. Різниця між статистичним та причинним зв'язком.
13. Вивчення статистичних зв'язків на основі причинних схем як основне завдання причинного аналізу.
14. Поняття допоміжної теорії вимірювань Блейлока.
15. Причинний аналіз як концептуальний підхід до вивчення соціальних явищ.
16. Проблема формалізації завдання вивчення причинно-наслідкових відносин в соціології.
17. Комплексне використання декількох методів вивчення зв'язків між ознаками для вирішення соціологічних задач (аналіз структури випадкової величини; факторний і дисперсійний аналіз; пошук детермінуючих поєднань значень предикторів).
18. Розв’язання практичних завдань.
**Література:** 1, 3, 4, 5, 6

## Тема 7. Завдання розпізнавання образів. Поняття автоматичної класифікації об'єктів (2 год.)

1. Класифікація як один із фундаментальних процесів у науці.
2. Загальне уявлення про завдання розпізнавання образів (синоніми: образ, клас, кластер, таксон; неоднозначність трактування термінів в літературі).
3. Виділення завдань: пошук класів, опис класів, визначення найбільш ефективної системи ознак.
4. Виділення задачі автоматичної класифікації об'єктів (синоніми: багатовимірна класифікація, розпізнавання образів без вчителя, кластерний аналіз, таксономія).
5. Ознаковий простір.

6. Задача класифікації як пошук згущення точок – моделей об'єктів в ознаковому просторі.
7. Роль наявності або відсутності навчальної вибірки.
8. Розв’язання практичних завдань.
**Література:** 1, 3, 4, 5, 6

## Тема 8. Проблема "стикування" змісту і формалізму при використанні алгоритмів класифікації (2 год.)

1. Специфіка рішення соціологічних завдань побудови типології за допомогою методів автоматичної класифікації.
2. Сенс протиставлення термінів "класифікація" і "типологія".
3. Виділення основних формальних елементів алгоритмів автоматичної класифікації, що вимагають стикування зі змістовними концепціями соціолога.
4. Підстава типології.
5. Роль апріорних уявлень дослідника про шуканих типах у виборі і реалізації алгоритму, інтерпретації результатів його застосування.
6. Розв’язання практичних завдань.
**Література:** 1, 3, 4, 5, 6

## Тема 9. Функції відстані між об'єктами (2 год.)

1. Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації.
2. Основні види гіпотез: компактності, зв'язності (безперервності), унімодального розподілу.
3. Приклади соціологічних завдань побудови типології, для яких була б розумна кожна гіпотеза.
4. Приклади алгоритмів, що шукають закономірності розташування точок у ознаковому просторі, що відповідають кожній з гіпотез: алгоритм Форель (гіпотеза компактності), алгоритм найближчого сусіда (гіпотеза зв'язності), алгоритм, заснований на виділенні локальних максимумів функції приналежності (гіпотеза унімодального розподілу ).
5. Роль функції належності у відповідних алгоритмах.
6. Змістовні уявлення соціолога про типи та умови вибору кроку розбиття при інтерпретації результатів.
7. Коригування результатів класифікації з метою забезпечення відповідності класифікації і типології.
8. Розв’язання практичних завдань.
**Література:** 1, 2, 3, 4, 5

## Тема 10. Основні види процедур класифікації. Відстані між класами (2 год.)

1. Виділення ієрархічних і неієрархічні алгоритмів класифікації.
2. Агломеративні та дівізімні алгоритми.
3. Оптимізація розбиття в сенсі максимізації заздалегідь обраного функціоналу якості як один з основних елементів формалізму в неієрархічні алгоритмах класифікації.
4. Основний змістовний сенс оптимізації. Сенс вимірювання близькості між класами в таких випадках.
5. Способи вимірювання сумарних оцінок близькості один до одного об'єктів усередині класів.
6. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5, 7

## Тема 11. Гіпотези про розташування об'єктів у ознаковому просторі

1. Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації.
2. Приклади соціологічних завдань побудови типології, для яких була б розумна кожна гіпотеза.
3. Загальне уявлення про розмиті класифікації.
4. Роль функції належності у відповідних алгоритмах.
5. Змістовні уявлення соціолога про типи та умови вибору кроку розбиття при інтерпретації результатів.
6. Розв’язання практичних завдань.

**Література:** 1, 3, 4

## Тема 12. Поняття інтерпретації вихідних даних і основні методологічні принципи використання методів аналізу даних в соціології (2 год.)

1. Інтерпретація вихідних даних як одне з основних ланок "стикування" соціології і математики.
2. Виділення методологічних принципів, дотримання яких є необхідним для того, щоб аналіз соціологічних даних був ефективний, не відводив соціолога в сторону від реальності: забезпечення певної однорідності вихідних даних; облік моделі, "закладеної" в кожному методі аналізу даних, при виборі алгоритму аналізу, два основні принципи інтерпретації результатів аналізу: необхідність її узгодження з інтерпретацією вихідних даних і заповнення при її здійсненні тих втрат, які мали місце при переході до формалізму; необхідність комплексного використання декількох методів для вирішення одного завдання і т. д.

3. Розв’язання практичних завдань.

**Література:** 1, 3, 4, 5

### Тема 13.Великі дані у соціології: нові дані, нова соціологія?і (2 год.)

1. Створення даних (DataGeneration/DataCapture)
2. Обслуговування даних (DataMaintenance)
3. Синтез даних (DataSynthesis)
4. Використання даних (DataUsage)
5. Публікація даних (DataPublication)
6. Архівація даних (DataArchival)
7. Знищення даних (DataPurging)
8. Розв’язання практичних завдань.

**Література:** 4, 8, 9

### Тема 14. Великі дані. Системи керування великими даними (2 год.)

1. Розподілені файлові системи
2. Розподілені фреймворки
3. Бенчмаркінг
4. Серверне програмування
5. Планування
6. Системи розгортання
7. Розв’язання практичних завдань.

**Література:** 4, 8, 9

### Тема 15. Програмні платформи та системи для Великих даних (2 год.)

1. Системи керування потоками даних
2. Системи зберігання Великих даних
3. Платформи Великих даних
4. Обробка даних у реальному часі
5. Системи керування Великими даними
6. Аналітичні платформи
7. Розв’язання практичних завдань.

**Література:** 4, 8, 9

**Тема 16. Машинне навчання за допомогою бібліотеки Scikit-learn. (2 год.)**

1. Кроки типового практичного сценарію машинного навчання.
2. Завантаження набору даних. Дослідження даних за допомогою Pandas. Візуалізація ознак за допомогою Matplotlib.
3. Налаштування параметрів моделі та оцінка її точності.
4. Функціонал бібліотеки Scikit-Learn. Класифікація за допомогою К-сусідів.
5. Лінійні моделі для регресії та класифікації (модель лінійної регресії, логістична регресія, та ін).
6. Дерева рішень та випадковий ліс.
7. Основи нейронних мереж.
8. Алгоритми кластеризації (кластеризація методом К-середніх, ієрархічна кластеризація, та ін).
9. Розв'язання практичних завдань.

**Література:** 4, 8, 9


Основна література

1. Горбачик А.П., Сальнікова С.А. Аналіз даних соціологічних дослідженьзасобами SPSS: Навч. посіб.- Луцьк, 2008. – 164 с. IBM SPSS 20 інструкція користувача// https://www.xn--80aaexjatkpdggghih8b1a2yhv.com.ua/ibm/spss-20/%D1%96%D0%BD%D1%81%D1%82%D1%80%D1%83%D0%BA%D1%86%D1%96%D1%8F-%D0%BA%D0%BE%D1%80%D0%B8%D1%81%D1%82%D1%83%D0%B2%D0%B0%D1%87%D0%B0.
2. Паніотто В.І., Максименко В. С., Харченко Н.М. Статистичний аналізсоціологічних даних. - Київ, 2004. – 270 с. Литвин В.В. Аналіз даних та знань: підручник/ В.В. Литвин, В.В. Пасічник, Ю.В. Нікольський.- Л.: Магнолія, 2020.- 276с. (базовий підручник).

Допоміжна література

3. Лупан І.В., Авраменко О.В., АкбашК.С.Комп'ютерні статистичні пакети: навчально-методичнийпосібник. - 2-е вид. - Кіровоград: 'КОД". 2015. - 230 с. - http://dspace.cuspu.edu.ua/jspui/bitstream/123456789.
MakingSenseofMultivariateData Analysis//https://us.sagepub.com/en-us/nam/book/making-sense-multivariate-data-analysis
4. Бахрушин В.Є.Методи аналізу даних: навчальний посібник для студентів В.Є. Бахрушин. - Запоріжжя : КПУ, 2011. - 26В с. - http://web.kpi.kharkov.ua/auts/wp-content/uploads/sites/67/2017/02/DAMAP_Ivashko_posobie2.pdf
5. Інтелектуальний аналіз даних: практикум/ М.Т. Фісун, І.О. Кравець, П.П. Казмірчук.- Л.: Новий Світ-2000, 2020.- 162с. Гладун А.Я., Рогушина Ю. В. DataMining: пошук знань в даних. Київ. ТОВ «ВД «АДЕФ- Україна», 2016. — 452 с..

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ**
**«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Кафедра соціології і публічного управління
(назва кафедри, яка забезпечує викладання дисципліни)

# ІНДИВІДУАЛЬНІ ЗАВДАННЯ З НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

# МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ
( назва навчальної дисципліни)

рівень вищої освіти другий (магістерський)
перший (бакалаврський) / другий (магістерський)

галузь знань 05 Соціальні та поведінкові науки
(шифр і назва)

спеціальність 054 Соціологія
(шифр і назва )

освітня програма Соціологічне забезпечення економічної діяльності
(назви освітніх програм спеціальностей )

вид дисципліни професійна підготовка; обов'язкова
(загальна підготовка / професійна підготовка; обов'язкова/вибіркова)

форма навчання денна
(денна / заочна/дистанційна)

Харків – 2024 рік

## Проектна робота

(вид індивідуального завдання)

| № з/п | | Терміни виконання (на якому тижні) |
|---|---|---|
| **1** | **Розробка індивідуального проекту щодо багатомірного шкалірування об'єктів за темою дипломної роботи** | 7-9 тижні |

## Реферат

| № з/п | Теми рефератів | Терміни виконання (на якому тижні) |
|---|---|---|
| 1 | **Теми рефератів**<br>1. Когнітивний аналіз опитувального інструментарію.<br>2. Огляд пакетів статистичного аналізу.<br>3. Основи статистики в комп'ютерних пакетах.<br>4. Так-система. Сильні і слабкі сторони логічних пакетів.<br>5. Програмування в SPSS.<br>6. Математичні методи в маркетингу.<br>7. Математичні методи вивчення конфліктних ситуацій.<br>8. Контент-аналіз і його реалізація на прикладі.<br>9. Дослідження Інтернету. Контент-аналіз, статистика та інтерпретація.<br>10. Факторний аналіз у політичних дослідженнях.<br>11. Кластеризація ринків.<br>12. Кластеризація національної економіки.<br>13. Кластеризація ціннісних орієнтацій.<br>14. Документообіг в світі безпаперових технологій.<br>15. Вітчизняний ринок програм обробки соціологічних даних.<br>16. Використання інформаційних систем для соціології: можливості і проблеми.<br>17. Консалтингова підтримка діяльності підприємств. | 2-3<br>4-5 |

| | | |
|---|---|---|
| | 18. Бази знань та експертні системи.<br>19. Проблеми захисту інформації в автоматизованих системах.<br>20. Інноваційні напрями розвитку інформаційних технологій.<br>21. Використання кореляційно-регресійного аналізу для обробки соціологічних даних.<br>22. Застосування первинних і вторинних угруповань в аналізі соціологічних даних.<br>23. Багатовимірні угруповання в статистиці<br>24. Роль графіків в узагальненні та аналізі соціологічних даних.<br>25. Графічний метод у вивченні соціальної реальності.<br>26. Графічне зображення в узагальненні та аналізі статистичних даних.<br>27. Середні величини в статистиці, їх значення, види.<br>28. Застосування структурних середніх величин для аналізу соціальних явищ.<br>29. Роль показників варіації в оцінці достовірності даних проведених досліджень.<br>30. Використання різних методик розрахунку показників варіації<br>31. Вибіркове спостереження, як основний метод проведення статистичного дослідження: його етапи, властивості, переваги та недоліки.<br>32. Предмет аналізу даних і витоки формування його методології.<br>33. Модель вивчення властивості об'єкту<br>34. Типи емпіричних даних<br>35. Шкалювання та кодування в процесі вимірювання.<br>36. Індекси при зборі та аналізі даних.<br>37. Специфічні прийоми вимірювання соціальної установки<br>38. Аналіз взаємозв'язку ознак<br>39. Заходи зв'язку, засновані на поняттях «статистична залежність» і «детермінації».<br>**40.** Заходи зв'язку засновані на моделі прогнозу | 6-7<br><br>8-9 |

# МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIGDATA В СОЦІОЛОГІЇ

**методичні вказівки з написання курсової роботи для студентів, що навчаються за напрямом «Соціологія»**

## ПЕРЕДМОВА

При підготовці фахівців у ЗВО важливе місце посідає навчально-дослідницька робота студентів, в якій особливе місце надається виконанню курсової роботи. В навчальному плані підготовки соціологів за рівнем «Магістр» передбачає написання курсової роботи з курсу «Методибагатовимірногоаналізу та BigData в соціології»на 1 курсі.Предметом вивчення курсу є оволодіння методамибагатовимірного аналізу даних, обробка даних соціологічних досліджень за допомогою програми SPSS.

Курсова робота повинна відповідати кваліфікаційним вимогам щодо змісту та оформлення. Слід пам'ятати, що науковий зміст курсової роботи завжди несе на собі печатку творчої індивідуальності автора, в той час як організація її підготовки підпорядковується загальному порядку, а оформлення – діючим стандартам. Виходячи з цього, рекомендації щодо наукового змісту робіт слід сприймати як консультативні, в той час як відомості про організацію підготовки роботи та правила її оформлення носять обов'язковий, нормативний характер.

Курсова робота з «Методибагатовимірногоаналізу та BigData в соціології»ставить своєю метою:

• поглибленевивчанняоднієї з тем курсу, яка пов‘язана з використанням методів багатомірного аналізу;

• вивченння творів вчених середини XX – початку XXI століття щодо можливості застосування багатовимірних методів для аналізу соціологічних

даних, а також кола критичної наукової літератури;

• опрацьовання різноманітних методів багатомірного аналізу даних;

• практичне використання багатовимірних методів для аналізу соціологічних даних.

# СТРУКТУРА КУРСОВОЇ РОБОТИ

Отже, будь-яка курсова робота починається з вступу. Справжній вступ містить всі найважливіші елементи: він починається з постановки конкретної проблеми – проблеминаписання курсової роботи, звідки слід обґрунтування актуальності теми курсової роботи, короткого огляду літератури за темою.

Визначається об’єкт та предмет курсової. Потім формулюються мета та завдання, які будуть реалізовані у роботі.

На завершення слід привести важливе правило: вступ до наукової роботи, як і висновок, рекомендується писати після повного завершення основної частини. До того, як буде створено основну частину роботи, неможливо написати гарний вступ, так як автор ще не цілком опанував матеріалами за темою.

**У ході написання роботи студент повинний вирішити наступні задачі:**

- з'ясувати актуальність теми дослідження, запропонованої в курсовій роботі і розробленість її в наукових працях вітчизняних і закордонних авторів;
- визначити практичну значущість дослідження з даної теми і її зв'язок із процесами і явищами, що відбуваються в країні;
- виділити об’єкт, предмет, мету та завдання для роботи;
- строго й аргументовано викласти основні ідеї різних авторів, щодо до багатомірного на налізу даних, обрати певний підхід (чи декілька підходів) як подальшу методологічну та теоретичну базу власного аналізу проблеми;
- на основі зробленого аналізу в теоретичній частині роботи, описати один з багатомірних методів, його можливості при застосуванні аналізу соціологічних даних;
- провести збір та первинну обробку даних;
- застосувати комплекс багатомірних методів до аналізу отриманих даних, а саме факторний, багатовимірне шкалірування, кластерний та побудова дерев рішень;

- описати отримані результати з використанням таблиць та графіків;
- надати практичні рекомендації виходячи з результатів дослідження.

Робота носить дослідницький й аналітичний характер. В ході її написання студенту повинно використовувати різні джерела інформації, якими можуть бути:

- навчальна література з відповідних курсів;
- підручники, монографії, журнальна періодика що містить новітню інформацію за обраною проблемою;
- довідкова література (словники), що містить тлумачення основних понять;
- електронні засоби масової інформації – Інтернет.

Отже курсова робота складається з наступних складових частин:

- титульного листа;
- змісту;
- вступу;
- основної частини, що складається з окремих розділів, які можуть, в свою чергу, поділятися на пункти та підпункти;
- висновків;
- списку літератури;

Загальний обсяг бакалаврської роботи складає 15-20 сторінок (без додатків та списку літератури).

Кожна частина роботи починається з нової сторінки. Маються на увазі титульний лист, зміст, розділи, висновки, список джерел інформації, додатки.

Усі сторінки роботи повинні бути пронумеровані, починаючі з титульної (вона рахується, але номер на ній не ставиться). Номер сторінки ставиться у верхньому правому куті сторінки.

**Більш детально правила оформлення курсової роботи дивись далі у цих методичних рекомендаціях**

**Титульний лист**містіть назву міністерства, університету, кафедри, в межах якої виконується робота. Далі надається тема роботи, інформація щодо виконавця та керівника роботи. Наприкінці вказується місто та рік виконання (Приклад оформлення титульного листа надано у Додатку А цих методичних рекомендацій).

**Зміст**роботи включає перелік основних її частин із указівкою сторінок їхнього початку. Назва кожної частини записується з нового рядка з заголовної букви. Вступ, висновки, список джерел інформації і додаток не нумеруються. Основна частина складається з окремих розділів, кожний з який нумерується, має власну назву (яка пишеться без лапок) і записується з нового рядка з указівкою сторінки початку (Додаток Ж). Розділи повинні бути зв'язані між собою логічно. Їх повинно бути три.

**Вступ** повинен містити ряд обов’язкових елементів: обґрунтування теми, її актуальність в теоретичному та практичному плані, ступінь її розробленості в наукової літературі, чітко сформульовані об’єкт, предмет, цілі та завдання дипломної роботи. Вступ займає 2-3 сторінки.

Об’єктомкурсової роботи можуть виступати реальні соціальні процеси або явища.

Предметом роботи при цьому є багатомірні методи аналізу соціологічних даних щодо визначеного об’єкту курсової.

Метою роботи є той результат, який виконавець планує досягти в результаті здійснення роботи. Метою може бути визначення за допомогою багатомірних методів аналізу характеристик певного явища чи процесу.

Завдання роботи уточнюють мету, розкривають основні етапи її досягнення. Завдання роботи повинні бути як теоретичного, так і практичного плану. Завдань в роботі може бути 3-4 (вивчити особливості соціологічного дослідження об’єкту дослідження, описати можливості застосування сучасних методів багатомірного дослідження соціологічних даних, застосувати методи багатомірного дослідження для вивчення об’єкту, надати рекомендації)

**Основна частина** роботи може складатися з трьох розділів. У розділах послідовно розкривається тема роботи.

Кожний розділ має бути обсягом 5 – 7 сторінок. Перший розділ стислий огляд (опис) соціального об'єкту. Другий розділ опис одного з методів багатомірного аналізу – факторний, багатомірне шкалірування, кластерний аналіз, дерево рішень, дискретний аналіз. Третій розділ опис застосування всіх методів для вивчення емпіричних даних.

**Висновки** – обов'язкова частина курсової роботи – являють собою стисле викладення одержаних автором наукових результатів, які формулюються у вигляді окремих пунктів. Наприкінці висновків необхідно надати рекомендації за результатами дослідження. Обсяг висновків до 2 сторінок.

**Список джерел інформації** містить всі наукові праці, що розглядались в літературному огляді за темою курсової роботи (вимоги щодо оформлення бібліографії наведені в Додатку Б цих методичних рекомендацій). **Він містить не менш 10 основних наукових джерел за темою.** Обов'язкова присутність усіх видів джерел: навчальної літератури за курсами вітчизняних і закордонних авторів; наукових публікацій за проблемою, ідеї яких були використані в роботі. Джерел повинно бути достатньо для розкриття теми, а інформація, що міститься в них – новою і достовірною.

## ЗОВНІШНЄ ОФОРМЛЕННЯ КУРСОВОЇ РОБОТИ

Курсова робота повинна бути представлена державною мовою в друкованому та електронномуваріантах. Друкованийваріант повинен бути переплетений (м'якаабо тверда палітурка).

Курсовуроботу друкують за допомогою принтера на одному боціаркушабілогопаперу формату А4 (210 х 297 мм) через ***1,5 інтервали 14 шрифтом TimesNewRoman.***Текст повинен бути вирівняний***по ширині.***

Текст роботидрукують, залишаючи поля таких розмірів: ***ліве - 30 мм, праве - 15 мм, верхнє - 20 мм, нижнє - 20 мм.*** Шрифт друку повинен бути чітким, чорногокольорунежирним. Щільність тексту всюдиоднакова.

Першоюсторінкоюкурсової є титульнийаркуш, якийпідлягаєзагальнійнумераціїсторіноккурсової. ***На титульному аркуші номер сторінки не ставлять,*** на наступних - номер проставляють у ***правому верхньому куті сторінки без крапки в кінці.***

Друкарськіпомилки, описки, графічнінеточності, яківиявилисяпід час написанняроботи, можнавиправлятипідчищеннямабозафарбуваннямбілоюфарбою та нанесенням на тому ж місціабоміж рядками виправленого тексту друкарськимилітерами. Допускаєтьсянаявність не більшедвохвиправлень на однійсторінці.

***Оформлення тексту роботи***

Текст основноїчастинидипломноїроботиподіляють на розділи. Заголовки структурнихчастинкурсовоїроботи «ЗМІСТ», «ВСТУП», «ВИСНОВКИ», «СПИСОК ДЖЕРЕЛ ІНФОРМАЦІЇ» друкують великими літерамисиметричнодо наборута виділяютьнапівжирним шрифтом.

Номер розділуставлять до назвивідповідногорозділу, після номера крапку не ставлять, потімдрукують заголовок розділу. Назвурозділувиділяютьнапівжирним. Крапку в кінці заголовка не ставлять.

Кожнуструктурнучастинукурсовоїроботи треба починати з новоїсторінки.

***Оформленняілюстративногоматеріалу***

В курсових роботах можезастосовуватисярізноманітнийілюстративнийматеріал (рисунки, схеми, таблицітощо). Існуютьзагальноприйняті правила оформленняілюстративногоматеріалу, якимислідкеруватися при підготовцікурсовоїроботи.

***Ілюстрації.***Кількістьілюстрацій (рисунків, схем, графіківтощо) в курсовійроботівизначаєтьсязмістомостанньої та повинна бути достатньою для того, щобнадатитекстовіясності й конкретності.

Ілюстраціїнеобхідноподавати в курсовоїроботібезпосередньопісля тексту, де вони згаданівперше, або на наступнійсторінці.

Ілюстраціїпозначають словом «Рисунок» і нумеруютьпослідовно в межах розділу за виняткомілюстрацій, наведених у додатках. Номер ілюстрації повинен складатися з номера розділу та порядкового номера ілюстрації, міжякими ставиться крапка. Наприклад, Рисунок 1.2 (другий рисунок першогорозділу). Спочаткурозміщуютьілюстрацію, потімпід нею симетрично тексту пишутьїї номер та назву. Післяназви рисунка крапка не ставиться, названапівжирним не виділяється. Зверху та знизу рисунка залишаєтьсявільний рядок.

Якщо в розділікурсової подано одну ілюстрацію, то їїнумерують за загальними правилами. Ілюстраціїнумеруються в межах розділуокремовідтаблиць, тобто у ілюстрацій своя нумерація, а у таблиць своя.

Не вартооформлюватипосилання на ілюстрації як самостійніфрази, в якихлишеповторюється те, щоміститься у підписі. У тому місці, де викладається тема, пов'язана з ілюстрацією, і де читачеві треба вказати на неї, розміщуютьпосилання у виглядівиразу в круглих дужках «(рис. 1.2)» абозворот типу: «...як це видно з рис. 1.2» або «... як це показано на рис. 1.2».

***Таблиці***

Таблицінумеруютьпослідовно в межах розділу. Спочаткузліва, з абзацу розміщуютьнапис «Таблиця» іззазначеннямії номера. Номер таблиці повинен складатися з номера розділу та порядкового номера таблиці, міжякими ставиться крапка, наприклад: «Таблиця 1.2» (друга таблицяпершогорозділу).

Кожнатаблиця повинна мати заголовок, якийміститься рядом зі словом «Таблиця» безпосередньонад самою таблицею. Заголовок пишуть з великоїлітери в підбірдо тексту. Післяназвитаблицікрапку не ставлять.

Якщо в розділі одна таблиця, їїнумерують за загальними правилами. При перенесеннічастинитаблиці на іншусторінку з абзацу пишуть: «Закінченнятаблиці (номер)».

Зверху на знизутаблицірекомендуєтьсязалишати один вільний рядок.

Таблицю, залежновідїїрозміру, можнавміщуватипісля тексту, у якому вона згадується.

Кожнатаблиця повинна мати головку з заголовками граф та підзаголовками, боковину з заголовками рядків, рядки й графи. Кожен заголовок над графою стосуєтьсявсіхданихцієї графи, кожен заголовок рядка в боковику-всіхданихцього рядка.

Заголовок кожної графи в головцітаблицімусить бути по можливості коротким. Слідуникатиповторівтематичного заголовка в заголовках граф, одиницівимірузазначати у тематичному заголовку, виносити до узагальнюючихзаголовків слова, щоповторюються. Боковик, як і головка, потребуєлаконічності. Повторювані слова тут такожвиносять в об'єднувальні рубрики; загальні для всіхзаголовківбоковика слова розміщуютьу заголовку над ним.

Заголовки граф повинніпочинатися з великих літер, підзаголовки - з маленьких, якщо вони складаютьоднереченняіз заголовком, і з великих, якщо вони є самостійними.

Якщо в таблиці є текст, щоповторюється, якийскладається з одного слова, йогоможназамінитилапками, якщоповторюється текст двох і більшеслів, то при першомуйогоповторенні текст замінюють словами «тежсаме», а потім лапками. Ставити лапки замість цифр, знаків, символів, щоповторюються, неможна. Якщов будь-якійграфітаблиціцифровіабоіншіданівідсутні, то на цьомумісцівставлять риску.

Таблицюрозміщують в текстіпісляпершоїзгадки про неї, а при переносітаблиці до наступноїсторінки головку таблиціслідповторити. Якщо головка таблицінадтогроміздка, їїможна не повторювати на наступнійсторінці, а, пронумерувати графи, повторитицюнумерацію. Розділяти головки таблиць за діагоналлюнеможна.

***Загальні правила цитування та посилання на використаніджерела***

Оскількиосновний текст роботистворюється з опорою на різноманітнілітературніджерела, то особливуувагувартоприділитипосиланням на використанулітературу. Такіпосиланнядаютьзмогувідшукатидокументи, перевіритиїхдостовірність, допомагаютьз'ясуватийогозміст, мову тексту, обсяг. Посиланнямусятьсупроводжувативсіцитати, а такожідеї, що автор запозичив з працьвчених.

Посилання в текстіроботи на джереласлідоформлювати в квадратних дужках. В цихдукахспочаткувказується номер джерела на яке посилаються, щовідповідаєйого порядковому номеру у спискуджерелінформаціїкурсової, а потім номер сторінки (сторінок), на якихміститьсяідея, щоцитується.

Якщо автор використавлишезагальну думку, висновок, теорію, класифікацію, періодизацію і виклавїїсвоїми словами, то відповіднепосиланняможе бути поставленепісля абзацу і міститилише номер науковогоджерелаіззагального списку літератури.

Якщомаємісценепрямецитування, тобтопереказ, виклад думок іншихавторівсвоїми словами, то в квадратних скобках достатньовказати номер джерела, аботжерелу спискулітератури.

Якщо в роботімаємісцедослівнецитування, то необхіднодотримуватись таких вимог:

• текст цитатипочинається і закінчується лапками та наводиться в тійграматичнійформі, в якійвінподаний у джерелі, зізбереженнямособливостейавторськогонаписання.

• цитування повинно бути повним, без довільногоскороченняавторського тексту та без перекручень думок автора. Пропуск слів, речень, абзаців при

цитуваннідопускається без перекрученняавторського тексту і позначаєтьсятрьомакрапками. Вони ставляться у будь-якомумісціцитати (на початку, всередині, наприкінці). Якщо перед випущеним текстом або за ним стояв розділовий знак, то він не зберігається;

• кожна цитата обов'язковосупроводжуєтьсяпосиланням на джерело.

***Оформлення списку джерелінформації.***

Список джерелінформаціїміститьбібліографічні описи використанихджерел. Вінмістить не меншніж1***0 основних джерел інформації*** у наступному порядку: на початку списку проводятьсяпершоджерела, потімроботи, щонадрукованіукраїнською і російськоюмовами, а за ними іноземними (кожнагрупа у алфавітному порядку). Роботи одного автора можутьрозташовуватись у хронологічному порядку або за алфавітом (по назвах).

Відомості про книги повиннівключати: прізвище та ініціали автора, назву книги, місцевидання, видавництво і ріквидання, обсяг книги в сторінках. Прізвище автора слідвказувати у називномувідмінку. Якщо книга написана двома і більше авторами, то їхпрізвища та ініціаливказують у тій же послідовності, в якій вони надруковані в книзі. Назва книги приводиться у тому вигляді, в якому вона міститься на титульному листі. Найменуваннямісцявиданнянеобхіднодаватиповністю в називномувідмінку.

Прикладиоформленнябібліографічногоописуу списку літературинадано в ДодаткуБ.

## КРИТЕРІЇ ОЦІНКИ КУРСОВОЇ РОБОТИ

Нижчінаводятьсякритерії, за якимивиставляєтьсяпідсумковаоцінка за курсову роботу. Комісія разом з керівникомкурсовоїроботивизначаютькількіснезначення кожного зкритеріїв за 10-бальною шкалою.

Критерії оцінки курсової роботи

| | Критерію оцінки | Бал | Питома |
|---|---|---|---|
| | Оформлення роботи: відповідність до стандартів і вимог кафедри | 1-10 | 25% |
| | Змістовність, логічність та структурованість роботи | 1-10 | 50% |
| | Вміння стисло, логічно і повно доповісти про результати роботи та відповісти на питання членів комісії | 1-10 | 25% |

Отримані студентом балидодаються і перераховуються у більшзвичну систему оцінок (табл.3.2).

Система переводу оцінок

| Підсумкова оцінка | Оцінка за шкалою ECTS | Оцінка за національною шкалою |
|---|---|---|
| 10 | A | 5 |
| 9 | A | 5 |
| 8 | B | 4 |
| 7 | C | 4 |
| 6 | D | 3 |
| 5 | E | 3 |
| 0-4 | FX | 2 |

## ГРАФІК НАПИСАННЯ КУРСОВОЇ РОБОТИ

| Види робіт | Термін виконання |
|---|---|
| 1. Визначення теми курсової роботи | Перший тиждень березня |
| 2. Попередній аналіз літератури, складання плану роботи та його обговорення з керівником | До кінцяберезня |
| 3. Детальний аналіз літератури та написання змісту, вступу та першого розділу курсової роботи | квітень |
| 6. Написання останніх розділів курсової роботи, висновків, оформлення списку джерел інформації | травень |
| 7. Написання висновків, оформлення списку джерел інформації | травень |
| 9. Надання готової курсової роботи на кафедру для реєстрації | Кінцевий термін 25травня |
| 10. Захист курсової роботи | Кінець травня |

**Додаток А**
**Приклад оформлення титульного листа**

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**
**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ**
**«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Факультет соціально-гуманітарних технологій

Кафедра соціології і публічного управління

Курсова робота з дисципліни «Методибагатовимірногоаналізу та BigData в соціології» на тему:

**ОСОБЛИВОСТІ ЗАСТОСУВАННЯ БАГАТОМІРНИХ МЕТОДІВ АНАЛІЗУ ДО ВИВЧЕННЯ …**

**Виконав:**
студент групи СГТ М 57
Петренко Г.Г.

**Науковий керівник:**
професор кафедри соціології
і публічного управління,
доктор соціологічних наук
Бірюкова М.В.

ХАРКІВ 2023

ДодатокБ

**Приклад оформленнябібліографічногоописуу спискулітератури, який наводиться у курсовійроботі**

Характеристика джерела. Приклад оформлення

КНИГИ: Один автор

1. Василій Великий. Гомілії / ВасилійВеликий ; [пер. з давньогрец. Л. Звонська]. — Львів :Свічадо, 2006. — 307 с. — (Джерелахристиянського Сходу. Золотийвік патристики IV—V ст. ; № 14).

2. КоренівськийД. Г. Дестабілізуючийефектпараметричногобілого шуму в неперервних та дискретнихдинамічних системах / КоренівськийД. Г. — К. :Ін-т математики, 2006. — 111 с. — (Математика та їїзастосування) (Праці / Ін-т математики НАН України ; т. 59).

3. Матюх Н. Д. Щодорожчесрібла-золота / Наталія ДмитрівнаМатюх. — К. :Асамблеяділ. кіл : Ін-т соц. іміджмейкінгу, 2006. — 311 с. — (ЮвеліриУкраїни ; т. 1).

4. Шкляр В. Елементал : [роман] / Василь Шкляр. — Львів :Кальварія, 2005. — 196, [1] с. — (Першотвір).

Два автори

1. Матяш І. Б. ДіяльністьНадзвичайноїдипломатичноїмісії УНР в Угорщині :історія, спогади, арх. док. / І. Матяш, Ю. Мушка. — К. :Києво-Могилян. акад., 2005. — 397, [1] с. — (Бібліотеканауковогощорічника "Україна дипломатична" ;вип. 1).

2. Ромовська З. В. СімейнезаконодавствоУкраїни / З. В. Ромовська, Ю. В. Черняк. — К. : Прецедент, 2006. — 93 с. — (Юридичнабібліотека. Бібліотека адвоката) (Матеріали до складанняквалiфікаційнихіспитів для отриманняСвідоцтва про право на заняттяадвокатськоюдіяльністю ;вип. 11).

3. Суберляк О. В. Технологіяпереробкиполімерних та композиційнихматеріалів :підруч. [для студ. вищ. навч. закл.] / О. В.

Суберляк, П. І. Баштанник. — Львів : Растр-7, 2007. — 375 с.

Три автори

1. Акофф Р. Л. Идеализированное проектирование: как предотвратить завтрашний кризис сегодня. Создание будущего организации / Акофф Р. Л., Магидсон Д., Эддисон Г. Д. ; пер. с англ. Ф. П. Тарасенко. — Днепропетровск : Баланс Бизнес Букс, 2007. — XLIII, 265 с.

Чотириавтори

1. Методика нормуванняресурсів для виробництвапродукціїрослинництва / [ ВітвіцькийВ. В., Кисляченко М. Ф., Лобастов І. В., Нечипорук А. А.]. — К.:

НДІ "Украгропромпродуктивність", 2006. — 106 с. — (БібліотекаспеціалістаАПК. Економічнінормативи).

2. Механізаціяпереробноїгалузіагропромисловогокомплексу : [підруч. для учнів проф.-техн. навч. закл.] / О. В. Гвоздєв, Ф. Ю. Ялпачик, Ю. П. Рогач, М. М. Сердюк. — К. :Вищаосвіта, 2006. — 478, [1] с. — (ПТО: Професійно- технічнаосвіта).

П’ять і більшеавторів

1. Психология менеджмента / [ Власов П. К., Липницкий А. В., Лущихина И. М. и др.] ; под ред. Г. С. Никифорова. — [3-е изд.]. — Х. :Гуманитар. центр, 2007. — 510 с.

2. Формування здорового способу життя молоді : навч.-метод. посіб. для працівників соц. служб для сім’ї, дітей та молоді / [Т. В. Бондар, О. Г. Карпенко, Д. М. Дикова-Фаворська та ін.]. — К. : Укр. ін-т соц. дослідж., — 115 с. — (Серія "Формування здорового способу життя молоді" : у 14 кн., кн. 13).

Без автора

1. Історія Свято-МихайлівськогоЗолотоверхого монастиря/ [авт. тексту В. Клос]. — К. :Грані-Т, 2007. — 119 с. — (Гранісвіту).

2. Воскресіннямертвих :українськабарокова драма : антологія / [упорядкув., ст., пер. і прим. В. О. Шевчук]. — К. : Грамота, 2007. — 638, [1]

с.

3. Тілочиособистість? Жіночатілесність у вибраніймалійукраїнськійпрозі та графіцікінця XIX — початку XX століття : [антологія / упоряд.: Л. Таран,Лагутенко]. — К. :Грані-Т, 2007. — 190, [1] с.

4. Проблемитипологічної та квантитативноїлексикології : [зб.наук.праць / наук. ред. Каліущенко В. та ін.]. — Чернівці : Рута, 2007. — 310 с.

Багатотомний документ

1. ІсторіяНаціональноїакадеміїнаукУкраїни, 1941—1945 / [упоряд. Л. М. Яременкотаін.]. — К. :Нац. б-каУкраїниім. В. І. Вернадського, 2007— .— (Джерела з історіїнауки в Україні).Ч. 2 :Додатки — 2007. — 573, [1] с.

2. Межгосударственныестандарты :каталог в 6 т. / [сост. Ковалева И. В., Рубцова Е. Ю. ; ред. Иванов В. Л.]. — Львов : НТЦ "Леонорм-Стандарт", 2005— .— (Серия "Нормативнаябазапредприятия"). Т. 1. — 2005. — 277 с.

3. Дарова А. Т. НеисповедимыпутиГосподни... : (Дочьвраганарода) : трилогия / А. Дарова. — Одесса :Астропринт, 2006— .— (Сочинения : в 8 кн. / А. Дарова ; кн. 4).

4. Кучерявенко Н. П. Курсналоговогоправа :Особеннаячасть : в 6 т. / Н. П. Кучерявенко. — Х. Право, 2002— Т. 4: Косвенныеналоги. — 2007. — 534 с.

5. Реабілітованііісторією. Житомирськаобласть : [у 7 т.]. — Житомир :Полісся, 2006— .— (Науково-документальнасеріякниг "Реабілітованііісторією" : у 27 т. / голов. редкол.: Тронько П. Т. (голова) [таін.]). Кн. 1 / [обл. редкол.: Синявська І. М. (голова) таін.]. — 2006. — 721, [2] с.

6. Бондаренко В. Г. Теоріяймовірностей і математичнастатистика. Ч.1 / В. Г. Бондаренко, І. Ю. Канівська, С. М. Парамонова. — К. : НТУУ "КПІ", 2006. — 125 с.

Матеріаликонференцій, з’їздів

1. Економіка, менеджмент, освіта в

системіреформуванняагропромисловогокомплексу :матеріалиВсеукр. конф. молодихучених-аграрників ["Молодь України і аграрна реформа"], (Харків, 11—13 жовт. 2000 р.) / М-во аграр. політики, Харк. держ. аграр. ун-т ім. В. В. Докучаєва. — Х. :Харк. держ. аграр. ун-т ім. В. В. Докучаєва, 2000. — 167 с.

2. Кібернетика в сучаснихекономічнихпроцесах :зб. текстіввиступів на республік. міжвуз. наук.-практ. конф. / ДержкомстатУкраїни, Ін-т статистики, обліку та аудиту. — К. : ІСОА, 2002. — 147 с.

3. Матеріали ІХ з'їздуАсоціаціїукраїнськихбанків, 30 червня 2000 р. інформ. бюл. — К. :Асоц. укр. банків, 2000. — 117 с. — (Спецвип.: 10 років АУБ).

4. Оцінка й обгрунтуванняпродовження ресурсу елементівконструкцій :праціконф., 6—9 черв. 2000 р., Київ. Т. 2 / відп. Ред. В. Т. Трощенко. — К. :НАН України, Ін-т пробл. міцності, 2000. — С. 559—956, XIII, [2] с. — (Ресурс 2000).

5. Проблемиобчислювальноїмеханіки і міцностіконструкцій :зб. наук. праць / наук. ред. В. І. Моссаковський. — Дніпропетровськ :Навч. кн., 1999. — 215 с.

6. Ризикологія в економіці та підприємництві :зб. наук. праць за матеріаламиміжнар. наук.-практ. конф., 27-28 берез. 2001 р. / М-во освіти і науки України, Держподатк. адмін. України [та ін.]. — К. : КНЕУ : Акад. ДПС України, 2001. — 452 с.

ПРЕПРИНТИ

1. ШиляевБ. А. Расчеты параметров радиационного повреждения материалов нейтронами источника ННЦ ХФТИ/ANL USA с подкритической сборкой, управляемой ускорителем электронов / Шиляев Б. А., Воеводин В. Н. — Х. ННЦ ХФТИ, 2006. — 19 с. — (Препринт / НАН Украины, Нац. науч. центр "Харьк. физ.-техн. ин-т" ; ХФТИ 2006-4).

2. Панасюк М. І. Про точністьвизначенняактивностітвердихрадіоактивнихвідходів гамма-методами / Панасюк М. І., Скорбун А. Д., Сплошной Б. М. — Чорнобиль :Ін-т пробл.

безпеки АЕС НАН України, 2006. — 7, [1] с. — (Препринт / НАН України, Ін-т пробл. безпекиАЕС ; 06-1).

ДЕПОНОВАНІ НАУКОВІ ПРАТИ

1. Тимошенко З. І. Болонськийпроцес в дії:словник-довідникоснов. термінів і понять з орг. навч. процесуу вищ. навч. закл. / З. І. Тимошенко, О.І.Тимошенко. — К. :Європ. ун-т, 2007. — 57 с.

2. Українсько-німецькийтематичнийсловник [уклад. Н. Яцко та ін.]. — К. : Карпенко, 2007. — 219 с.

3. ЄвропейськийСоюз :словник-довідник/ [ред.-упоряд. М. Марченко]. — 2- ге вид., оновл. — К. :К.І.С., 2006. — 138 с.

ЗАКОНОДАВЧІ ТА НОРМАТИВНІ ДОКУМЕНТИ

1. Кримінально-процесуальний кодекс України : за станом на 1 груд. 2005 р. / Верховна Рада України. — Офіц. вид. — К. :Парлам. вид-во, 2006. — 207 с.(Бібліотекаофіційнихвидань).

2. Медична статистика статистика :зб. нормат. док. / упоряд. та голов. ред. В. М. Заболотько. — К. : МНІАЦ мед. статистики : Медінформ, 2006. — 459 с.

3. Експлуатація, порядок і терміниперевіркизапобіжнихпристроїв посудин, апаратів і трубопроводівтепловихелектростанцій : СОУ-Н ЕЕ 39.501:2007.

БІБЛІОГРАФІЧНІ ПОКАЗЧИКИ

1. Куц О. С. Бібліографічнийпокажчикта анотаційкандидатськихдисертацій, захищених у спеціалізованійвченійрадіЛьвівського державного університетуфізичноїкультури у 2006 році / О. Куц, О. Вацеба. — Львів : Укр. технології, 2007. — 74 с.

2. Систематизованийпокажчикматеріалів з кримінального права, опублікованих у ВісникуКонституційного Суду України за 1997—2005 роки / [уклад. Кирись Б. О., ПотланьО. С.]. — Львів :Львів. держ. ун-т внутр. справ,

2006. — 11 с. — (Серія: Бібліографічнідовідники; вип. 2).

ДИСЕРТАЦІЇ

1. Петров П.П. Активністьмолодихзіроксонячноїмаси: дис. ... доктора фіз. - мат. наук : 01.03.02 / Петров Петро Петрович. - К., 2005. - 276 с.

АВТОРЕФЕРАТИ ДИСЕРТАЦІЙ

1. Новосад І.Я. Технологічнезабезпеченнявиготовленнясекційробочихорганівгнучкихгвинтовихконвеєрів :автореф. дис. на здобуття наук. ступеняканд. техн. наук : спец. 05.02.08 „Технологіямашинобудування” / І. Я. Новосад. — Тернопіль, 2007. — 20, [1] с.

2. Нгуен ШіДанг. Моделювання і прогнозуваннямакроекономічнихпоказників в системіпідтримкиприйняттярішеньуправліннядержавнимифінансами :автореф. дис. на здобуття наук. ступеняканд. техн. наук : спец. 05.13.06 „Автоматиз. системи упр. та прогрес. інформ. технології” / Нгуен ШіДанг. — К., 2007. — 20 с.

АВТОРСЬКІ СВІДОЦТВА

1. А. с. 1007970 СССР, МКИ3 В 25 J 15/00. Устройство для захвата неориентированных деталей типа валов / В. С. Ваулин, В. Г. Кемайкин(СССР). - № 3360585/25-08 ;заявл. 23.11.81 ;опубл. 30.03.83, Бюл. № 12. Патенти 1. Пат. 2187888 Российская Федерация, МПК7 Н 04 В 1/38, Н 04 J 13/00. Приемопередающее устройство / Чугаева В.И.; заявитель и патентообладатель Воронеж. науч.-исслед. ин-т связи. - № 2000131736/09 ;заявл. 18.12.00 ;опубл. 20.08.02, Бюл. № 23 (II ч.).

ЧАСТИНАКНИГИ, ПЕРІОДИЧНОГО, ПРОДОВЖУВАНОГО ВИДАННЯ

1. Козіна Ж. Л. Теоретичніоснови і результати практичного застосування системного аналізу в науковихдослідженнях в

областіспортивнихігор / Ж. Л. Козіна// Теорія та методика фізичноговиховання. — 2007. — № 6. — С. 15—18, 35—38.

2. Гранчак Т. Інформаційно-аналітичніструктурибібліотек в умовахдемократичнихперетворень / ТетянаГранчак, ВалерійГоровий // Бібліотечнийвісник. — 2006. — № 6. — С. 14—17.

3. Валькман Ю. Р. Моделирование НЕ-факторов — основа интеллектуализации компьютерных технологий / Ю. Р. Валькман, В. С. Быков, А. Ю. Рыхальский // Системнідослідженнята інформаційнітехнології. — 2007. — № 1. — С. 39—61.

4. МаШуінПроблемипсихологічноїпідготовкив системіфізкультурноїосвіти / МаШуін // Теорія та методика фізичноговиховання. — 2007. — № 5. — С. 12—14.

5. РегіональніособливостісмертностінаселенняУкраїни / Л. А. Чепелевська, Р. О. Моісеєнко, Г. І. Баторшина [та ін.] // Вісниксоціальноїгігієни та організаціїохорониздоров'яУкраїни. — 2007. — № 1. — С. 25—29.

6. ВаловаІ. Новіпринципи угоди Базель II / І. Валова; пер. з англ. Н. М. Середи // Банки та банківськісистеми. — 2007. — Т. 2, № 2. — С. 13—20.

7. Зеров М. ПоетичнадіяльністьКуліша // Українськеписьменство XIX ст. ВідКуліша до Винниченка : (нариси з новітнього укр., письменства) : статті / Микола Зеров. — Дрогобич, 2007. — С. 245—291.

8. Третьяк В. В. Возможности использования баз знаний для проектирования технологии взрывной штамповки / В. В. Третьяк, С. А. Стадник, Н. В. Калайтан // Современное состояние использования импульсных источников энергии в промышленности : междунар. науч.-техн. конф., 3-5 окт. 2007 г. : тезисы докл. — Х., 2007. — С. 33.

9. Чорний Д. Міськесамоврядування: тягаріпроблем, принадицивілізації/ Д. М. Чорний // По лівийбікДніпра: проблемимодернізаціїмістУкраїни:(кінецьXIX—початок XX ст. / Д. М. Чорний. — Х., 2007. — Розд. 3. — С. 137—202.

ЕЛЕКТРОННІ РЕСУРСИ

1. Богомольний Б. Р. Медицина екстремальнихситуацій [Електроннийресурс] ] : навч. посіб. для студ. мед. вузів III—IV рівніакредитації / Б. Р. Богомольний, В. В. Кононенко, П. М. Чуєв. — 80 Min / 700 MB. — Одеса : Одес. мед. ун-т, 2003. — (Бібліотека студента-медика) — 1 електрон. опт. диск (CD-ROM) ; 12 см. — Систем. вимоги: Pentium ; 32 Mb RAM ; Windows 95, 98, 2000, XP ; MS Word 97-2000.— Назва з контейнера.

2. Розподілнаселеннянайбільшчисленнихнаціональностей за статтю та віком, шлюбним станом, мовнимиознаками та рівнемосвіти [Електронний ресурс] : за данимиВсеукр. переписунаселення 2001 р. / Держ. ком. статистики України ; ред. О. Г. Осауленко. — К. :CD-вид-во "Інфодиск", 2004. — 1 електрон. опт. диск (CD-ROM) :кольор. ; 12 см. — (Всеукр. переписнаселення, 2001). — Систем. вимоги: Pentium-266 ; 32 MbRAM; CD-ROMWindows 98/2000/NT/XP. — Назва з титул. екрану.

Бібліотека і доступністьінформації у сучасномусвіті: електронніресурси в науці, культурі та освіті : (підсумки 10-ї Міжнар. конф. „Крим-2003”) [Електронний ресурс] / Л. Й. Костенко, А. О. Чекмарьов, А. Г. Бровкін, І. А. Павлуша // Бібліотечнийвісник — 2003. — № 4. — С. 43. — Режим доступу до журн. :ttp://www.nbuv.gov.ua/articles/2003/03klmko.htm

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ «ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Кафедра соціології і публічного управління
(назва кафедри, яка забезпечує викладання дисципліни)

# МЕТОДИ КОНТРОЛЮ ЗНАНЬ ЗНАВЧАЛЬНОЇ ДИСЦИПЛІНИ

## МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ

( назва навчальної дисципліни)

рівень вищої освіти другий (магістерський)
перший (бакалаврський) / другий (магістерський)

галузь знань 05 Соціальні та поведінкові науки
(шифр і назва)

спеціальність 054 Соціологія
(шифр і назва )

освітня програма Соціологічне забезпечення економічної діяльності
(назви освітніх програм спеціальностей )

вид дисципліни професійна підготовка; обов'язкова
(загальна підготовка / професійна підготовка; обов'язкова/вибіркова)

форма навчання денна
(денна / заочна/дистанційна)

Харків – 2024 рік

**Контрольні роботи** з курсу «МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ».

Контрольна робота полягає у самостійній роботі студента в поза аудиторний час над отриманими питаннями за темами курсу.

## Контрольна робота № 1

**Варіант № 1**
Проаналізуйте в чому полягає неодновимірність багатьох досліджуваних соціологом понять.
Визначте сутність переходу до простору довільної розмірності для успішного виконання завдання шкалірування.

**Варіант № 2**
Обґрунтуйте в чому полягає особлива роль простору сприйняття респондентами запропонованих їм об'єктів.
Визначте сутність моделі ідеальної точки в багатовимірному випадку.

**Варіант № 3**
Сформулюйте загальні принципи вивчення простору сприйняття як основного завдання БШ.
Обґрунтуйте в чому полягає особлива роль проблеми неметричного багатовимірного розгортання.

**Варіант № 4**
Проаналізуйте ідеї Кумбса щодо урахування можливості упорядкування відстаней між об'єктами.
Визначте сутність виду вихідних даних при багатовимірному шкалируванні.

**Варіант № 5**
Обґрунтуйте в чому полягає проблема формального визначення близькості.
Проаналізуйте в чому полягає роль соціолога при отриманні даних, вихідних для багатовимірного шкалірування, та інтерпретації його результатів.

**Варіант № 6**
Проаналізуйте процес створення функції відстані (аксіоматичне визначення).
Обґрунтуйте в чому полягають проблеми можливих способів одержання вихідних даних при багатовимірному шкалируванні.

**Варіант № 7**

Визначте особливості понять Евклідова відстань та Евклідовий простір.
Проаналізуйте особливості безпосереднього визначення близькості від респондентів, класифікація відповідних способів опитування; проблеми, що постають при такому способі збору даних.

**Варіант № 8**
Сформулюйте загальні характеристики неявного порівняння відстаней між близькістю, яке закладене у формулі функції стресу для метричного шкалірування.
Визначте особливості прикладів розрахунку матриці близькості на основі аналізу достатньо надійних даних іншого роду при багатовимірному шкалируванні.

**Варіант № 9**
Обґрунтуйте в чому полягає особлива роль поняття монотонної регресії, що використовується при розрахунку функції стресу для неметричного шкалірування.
Проаналізуйте особливості використання формальних та неформальних методів при інтерпретації результатів багатовимірного шкалірування.

**Варіант № 10**
Сформулюйте загальні характеристики важливості для соціології неметричного шкалірування.
Проаналізуйте значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей при багатовимірному шкалируванні.

**Варіант № 11**
Проведіть філософсько-соціологічний аналіз сутності формальних аспектів проблем розмірності евклідового простору і обертання, що визначають його осей координат.
Проаналізуйте в чому полягає особливості загального уявлення про моделювання частот таблиці спряженості при багатовимірному шкалируванні.

**Варіант № 12**
Сформулюйте загальні характеристики постановки завдання важливості врахування специфіки метрик окремих респондентів.
Визначте в чому полягає проблема мультиплікативних та адитивних моделей частот при багатовимірному шкалируванні.

**Варіант № 13**
Обґрунтуйте в чому полягає особлива роль визначення способу обліку метрик в індивідуальному БШ.
Проаналізуйте роль логарифмування мультиплікативної моделі при багатовимірному шкалируванні.

**Варіант № 14**
Визначте сутність виду вхідних і вихідних даних, функції стресу в індивідуальному БШ.
Обґрунтуйте в чому полягає можливість різного розуміння як сенсу розглянутих вкладів, так і того "середнього" рівня, з яким порівнюються спостерігаються частоти в процесі їх моделювання при багатовимірному шкалируванні.

**Варіант № 15**
Сформулюйте основні закономірності одномірного розгортання.
Проаналізуйте особливості використання поняття зв'язку між двома групами ознак при багатовимірному шкалируванні.

## Контрольна робота № 2

**Варіант № 1**
Визначте причини відхилення спостережуваних частот від їхніх середніх значень, тобто відмінності реального розподілу від рівномірного при багатовимірному шкалируванні.
Сформулюйте загальні характеристики важливості для соціолога вивчення відповідних зв'язків при багатовимірному шкалируванні.

**Варіант № 2**
Проаналізуйте особливості вибору моделів частот, що відповідають логлінейному аналізу.
Визначте принципову різницю між статистичним та причинним зв'язком в контексті багатовимірного шкалірування.

**Варіант № 3**
Обґрунтуйте в чому полягає особлива роль насиченої моделі при багатовимірному шкалируванні.
Визначте схеми обчислення ковариаций (кореляцій) між будь-якими двома ознаками на основі графа зв'язків в контексті багатовимірного шкалірування.

**Варіант № 4**

Сформулюйте основні вимоги до визначення мети переходу до логарифмів частот.
Сформулюйте загальні характеристики вивчення статистичних зв'язків на основі причинних схем в контексті багатовимірного шкалірування.

**Варіант № 5**
Визначте особливу роль гіпотези про взаємозв'язок ознак при багатовимірному шкалируванні.
Сформулюйте загальні характеристики обчислення структурних коефіцієнтів причинних схем в контексті багатовимірного шкалірування.

**Варіант № 6**
Проаналізуйте роль критерію "хі-квадрат" при використанні логлінейного аналізу.
Сформулюйте загальні характеристики основної теореми причинного аналізу в контексті багатовимірного шкалірування.

**Варіант № 7**
Проаналізуйте в чому полягає особливості розрахунку коефіцієнтів логлінейной моделі для двовимірного випадку.
Сформулюйте загальні характеристики поняття допоміжної теорії вимірювань Блейлока в контексті багатовимірного шкалірування.

**Варіант № 8**
Проаналізуйте особливості порівняння логлінейного аналізу з номінальним регресійним і дисперсійним аналізом, а також з методом послідовних розбивок.
Визначте роль і місце причинного аналізу як концептуального підходу до вивчення соціальних явищ в контексті багатовимірного шкалірування.

**Варіант № 9**
Сформулюйте загальні характеристики різного розуміння залежної ознаки: кількісна ознака в дисперсійному аналізі, кількісна або номінальна – в номінальному регресійному і частота, що стоїть в клітці багатовимірної таблиці спряженості, – в логлінейном аналізі.
Проаналізуйте проблеми формалізації завдання вивчення причинно-наслідкових відносин в соціології в контексті багатовимірного шкалірування.

**Варіант № 10**
Визначте принципові можливості пошуку поєднань значень предикторів: перевірка гіпотез про наявність багатовимірних зв'язків у логлінейном аналізі і можливість пошуку найбільш дієвих поєднань в методі послідовних розбивок і регресійному аналізі, заздалегідь заданий набір поєднань значень предикторів в дисперсійному аналізі.

Сформулюйте загальні характеристики агломеративних та дівізімних алгоритмів в контексті багатовимірного шкалірування

**Варіант № 11**
Визначте сутність поняття причини в соціології в контексті багатовимірного шкалірування.

Сформулюйте загальні характеристики функції відстані, які відмінні від евклідової: зважене евклидово, сіті-блок, Махаланобіса, Хеммінгово.

**Варіант № 12**
Визначте роль принципової неможливості повністю формалізувати поняття причини в контексті багатовимірного шкалірування.
Сформулюйте в чому полягає специфіка вирішення соціологічних завдань побудови типології за допомогою методів автоматичної класифікації в контексті багатовимірного шкалірування.

**Варіант № 13**
Проаналізуйте форми графу причинних зв'язків в контексті багатовимірного шкалірування.
Обґрунтуйте в чому полягає загальне уявлення про завдання розпізнавання образів в контексті багатовимірного шкалірування.

**Варіант № 14**
Проаналізуйте специфіку обчислення структурних коефіцієнтів в контексті багатовимірного шкалірування.
Визначте в чому полягає виділення завдань: пошук класів, опис класів, визначення найбільш ефективної системи ознак.

**Варіант № 15**
Визначте принципи побудови часткових коефіцієнтів кореляції і регресії в контексті багатовимірного шкалірування.

**Тестові завдання з курсу** «МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ»

## 1. Тестові завдання

1. Оберіть правильне твердження:
   а) нормальна випадкова величина ухиляється від свого середнього не більше, ніж на 2 кореня з дисперсії,
   б) нормальна випадкова величина ухиляється від свого середнього не більше, ніж на 3 кореня з дисперсії; в) нормальна випадкова величина

ухиляється від свого середнього не більше, ніж на 4 кореня з дисперсії.

2. Залежність виду Y = F (X) називається:
   а) лінійна кореляція; б) лінійна регресія; в) часткова кореляція.
3. Скільки залежних змінних може бути в рівнянні регресії:
   а) скільки завгодно; б) не більше 3; в) одна.
4. Не виконує завдання класифікації:
   а) кластерний аналіз,
   б) корреляційний аналіз;
   в) дискримінантний аналіз.
5. Чи можливо обчислити коефіцієнт регресії Y на X, якщо через відомо коефіцієнт кореляції і середньоквадратичне відхилення:
   а) не можливо, тому що необхідний показник дисперсії;
   б) можливо;
   в) залежить від виду аналізу.
6. До обмеження методу регресійного аналізу не належать:
   а) нормальність розподілу ознак;
   б) рівна кількість ознак змінних;
   в) змінні відмінні від нуля.
7. До обмежень методу факторного аналізу не належить:
   а) нормальність розподілу ознак;
   б) рівна кількість ознак змінних;
   в) рівність дисперсій.
8. До обмежень методу дисперсійного аналізу не належить:
   а) нормальність розподілу ознак;
   б) рівна кількість ознак змінних;
   в) рівність дисперсій.
9. Задачу прогнозування не виконує:
   а) дискримінантний аналіз;
   б) факторний аналіз;
   в) регресійний аналіз.
10. Для незалежних вибірок використовується:
   а) дисперсійний аналіз з повторними змінами;
   б) кореляційний аналіз;
   в) однофакторний дисперсійний аналіз.
11. Таке значення змінної менше якої мають рівно половина цих змінних, називається ...
   а) мода б) середнє
   в) медіана г) проміжне
12. Чому дорівнює кореляція Спірмена і Кендалла для двох змінних:

| № | X | Y |
|---|---|---|
| 1 | 32 | 2 |

а) 0

б) -1

| | | |
|---|---|---|
| 2 | 16 | 6 |
| 3 | 20 | 4 |
| 4 | 8 | 10 |
| 5 | 11 | 8 |

в) 1
г) 0,5

13.Чи може одне і те ж чисельне значення кореляції для різних вибірок мати різну статистичну значущість?

а) Так б) Ні

**2. Допишіть речення**

14.Дві вибірки є залежними, якщо …
15.Основний спосіб забезпечення репрезентативності вибірки відносно генеральної сукупності це …
16.Основна властивість вибірки, що визначає її якість – це …
17.Медіана як міра центральної тенденції не придатна для змінних в наступних шкалах …
18.Середнє як міра центральної тенденції не придатна для змінних в наступних шкалах …
19.Як співвідносяться дисперсії двох рядів чисел: 1) 5, 8, 10, 12, 11 і 2) 1, 4, 6, 8, 7 …
20.Гомогенність (рівність) дисперсій перевіряється перед …
21.Якщо при перевірці статистичної достовірності кореляції (при $\alpha = 0,05$) $p > 0.1$, то коректний висновок, що …
22.Для перевірки достовірності відмінності двох незалежних груп, члени яких ранжовані за ступенем вираженості «тривожності», використовують критерій
23.Для перевірки достовірності відмінності двох повторних вимірювань, члени яких ранжовані за ступенем вираженості «тривожності», використовують критерій …
24.Для перевірки достовірності відмінностей студентів 1 і 5 курсів за змінної «сімейний стан» (неодружений - ні) слід застосувати критерій …
25.Для перевірки достовірності відмінностей 2-х вибірок по перемінної "стать" (чоловік - дружин) слід застосувати критерій …
26.Для порівняння викладачів та студентів з «домінантності» (метрична шкала), слід застосувати критерій …
27.Якщо необхідно порівняти два повторних вимірювання кількісної змінної, що має помітні викиди, то застосовують критерій …
28.Для перевірки достовірності відмінності двох залежних вибірок по змінній, яка вимірюється в ранговій шкалі, застосовують критерій …

29.Для перевірки достовірності відмінності старших (1-я вибірка) і їхніх молодших (2-я вибірка) братів за рівнем домінантності, виміряної в метричній шкалі, застосовують критерій
30.Гіпотезу про взаємозв'язок читача (2 градації) і метричної змінних доцільно перевіряти за допомогою …
31.Статистична значимість поліпшення стану (рангова шкала) до і після терапії визначається за критерієм …
32.Гіпотезу про взаємозв'язок рангової і номинативной змінної, що має дві градації (напр., стать), доцільно вивчати за допомогою критерію …
33.Для перевірки гіпотези про відмінність 2 груп за ступенем індивідуальної мінливості (дисперсії) застосовують критерій …
34.Гіпотезу про взаємозв'язок метричної та номинативной змінної, що має дві градації (напр., стать), доцільно вивчати за допомогою критерію …
35.Гіпотезу про взаємозв'язок метричної та номинативной змінної, що має 5 градацій (наприклад, хобі), доцільно перевіряти за допомогою критерію …
36.Гіпотезу про взаємозв'язок порядкової і номинативной змінної, що має 4 градації (напр., посада), доцільно вивчати за допомогою критерію …
37.Гіпотезу про взаємозв'язок метричної і порядкової змінної, що має 15 градацій, доцільно вивчати за допомогою
38.Гіпотезу про взаємозв'язок 2-х кількісних змінних, що мають помітні викиди (асиметрії) доцільно перевіряти за допомогою
39.Для перевірки гіпотези про взаємозв'язок однієї метричної змінної та двох номінативних змінних доцільно застосовувати:
40.Для множинного порівняння середніх в рамках дисперсійного аналізу застосовують …
41.Метод множинного порівняння середніх який вимагає попереднього отримання статистично значимого результату дисперсійного аналізу це …
42.Багатовимірний дисперсійний аналіз призначений для вивчення впливу …
43.Модель багатовимірного дисперсійного аналізу включає …
44.Багатовимірний етап багатовимірного дисперсійного аналізу припускає-лага перевірку гіпотез …
45.Одновимірна етап багатовимірного дисперсійного аналізу проводиться для …
46.Дисперсійний аналіз з повторними вимірами дозволяє вивчать вплив на залежні змінні …
47.Різниця двох методів класифікації полягає в тому, що в першому задано число класів і приналежність деяких об'єктів до цих класів, а в другому – не задано ні те ні інше …
48.Подібність двох багатовимірних методів полягає в тому, що аналізуються кореляції між ознаками …
49.Частина дисперсії «залежною» змінної, обумовлена впливом «незалежних» змінних – це ...
50.Якщо незалежна змінна x в множині регресійному аналізі корелює з іншими незалежними змінними, то її внесок в дисперсію залежної змінної ...

51.Якщо незалежна змінна x в множині регресійному аналізі не корелює з іншими незалежними змінними, то її внесок в оцінку залежної змінної ...
52.Якщо в багатовимірному регресійному аналізі (y - залежна змінна, x1, x2 - незалежні змінні) r12 = 0,4; r1y = 0,8; r2y = - 0,5; β1 = 0,5; β2 = - 0, 2, то коефіцієнт множинної детермінації R2 дорівнює …
53.Метод «повної зв'язку» («далекого сусіда») в кластерному аналізі, в порівнянні з методом «одиночній зв'язку» («найближчого сусіда») дає ...
54.Метод «середньої зв'язку» (СЗ) в порівнянні з методами «далекого сусіда» (ДС) і «ближнього сусіда» (БС) зазвичай дозволяє отримати число кластерів …
55.Ієрархічний кластерний аналіз за (N - 1) кроків кластеризації (N - число об'єктів кластеризації) дає об'єднання …
56.Статистична значимість внеску кожної змінної на відміну класів визначається ...
57.Показником належності об'єкта до класу є ...
58.Основним заходом якості рішення в багатовимірному шкалировании є ...
59.Для попередньої оцінки числа факторів у факторному аналізі використовують ...
60.Величина факторної навантаження даної змінної з даного фактору свідчить ...

**Питання, що виносяться на екзамену з курсу «**МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIG DATA В СОЦІОЛОГІЇ»

---

1. Багатовимірне шкалірування: коло вирішуваних завдань.
2. Багатовимірне шкалірування: основні елементи формалізму (близькості, відстані, функція стресу).
3. Індивідуальне багатовимірне шкалірування: основні ідеї, мета використання в соціології, функція відстані.
4. Багатовимірне розгортання: основні ідеї, сенс вирішуються за його допомогою соціологічних завдань.
5. Багатовимірне шкалірування: проблеми формування вихідних даних і інтерпретації результатів.
6. Багатовимірне шкалірування: проблеми інтерпретації результатів.
7. Поняття багатовимірної зв'язку. Відносини переважання.
8. Логлінейний аналіз: мета використання в соціології, моделі частот.
9. Розрахунок параметрів логлінейной моделі для чотириклетинної таблиці спряженості. Зв'язок одержуваних величин з відносинами переважання.
10. Логлінейний аналіз: проблема формування гіпотез.
11. Порівняння можливостей логлінейного і номінального регресійного аналізу.
12. Порівняння можливостей логлінейного і дисперсійного аналізу.

13. Порівняння можливостей логлінейного аналізу та алгоритмів послідовних розбивок.
14. Канонічний аналіз: постановка завдання, канонічні кореляції.
15. Поняття канонічної кореляції як узагальнення множинного коефіцієнта кореляції.
16. Канонічний аналіз: моделі частот, використання при аналізі таблиць спряженості.
17. Канонічний аналіз: використання для побудови соціологічних індексів.
18. Оцифровка значень номінальних і порядкових ознак. Цілі використання відповідних методів. Подання про модель, що стоїть за кожним методом.
19. Канонічний аналіз як метод оцифровки.
20. Причинний аналіз: граф причинних зв'язків, структурні коефіцієнти, що координує шлях, його ефективність.
21. Причинний аналіз: обчислення коваріації (кореляції) будь-яких двох ознак на основі графа причинних зв'язків, шляхові коефіцієнти.
22. Зв'язок структурних коефіцієнтів з регресійним. Структурні рівняння.
23. Основна теорема колійного аналізу.
24. Роль латентних факторів у причинному аналізі.
25. Комплексне використання різних методів при аналізі структури взаємозв'язків ознак.
26. Загальне уявлення про завдання розпізнавання образів. Ознаковий простір. Поняття автоматичної класифікації об'єктів.
27. Сенс термінів "класифікація" і "типологія". Їх роль при вирішенні соціологічних завдань побудови типології об'єктів.
28. Роль функції відстані між об'єктами в процесі класифікації. Проблема її адекватності змістовному розуміння типу об'єктів.
29. Евклідова відстань. Виважена евклидова відстань. Відстань Хемінга.
30. Загальне уявлення про ієрархічних і неієрархічні алгоритмах класифікації.
31. Роль функції відстані між класами при реалізації алгоритмів класифікації. Її види.
32. Вибір форми шуканих класів при використанні методів класифікації. Гіпотези про розташування об'єктів у признаковом просторі.
33. Гіпотеза компактності. Алгоритм ФОРЕЛЬ
34. Гіпотеза зв'язності. Алгоритм найближчого сусіда.
35. Гіпотеза унімодального розподілу. Алгоритм, заснований на виділенні локальних максимумів функції приналежності.
36. Забезпечення відповідності класифікації і типології в процесі інтерпретації результатів класифікації.
37. Поняття інтерпретації даних. ЇЇ роль в соціології.
38. Принципи сполучення формалізму і змісту, зв'язку всіх етапів дослідження один з одним як основні методологічні принципи застосування методів аналізу даних в соціології. Приклади їх реалізації.

39. Забезпечення однорідності досліджуваної сукупності об'єктів як один з основних методологічних принципів застосування методів аналізу даних в соціології. Приклади його реалізації.
40. Методологічні принципи інтерпретації результатів застосування математичного методу для вирішення соціологічної завдання. Приклади їх реалізації.
41. Визначення даних. Філософський, юридичний підходи й життєвий цикл даних.
42.Поняття метаданих. Життєвий цикл метаданих
43. Специфікація системних вимог. Система метаданих
44. Розподілені файлові системи
45. Бенчмаркінг
46. Системи керування потоками даних
47. Системи зберігання Великих даних
48. Аналітичні платформи Види машинного навчання.
49. Основні бібліотеки машинного навчання Python (Scikit-learn, Keras, TensorFlow).
50. Функціонал бібліотеки Scikit-Learn.