**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ**

**«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

**Кафедра соціології та політології**

**КОНТЕНТ ЛЕКЦІЙ З НАВЧАЛЬНОЇ ДИСЦИПЛІНИ**

**«МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ДАНИХ В СОЦІОЛОГІЇ»**

2017

**Тема 1.1. Основні елементи формалізму**

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.

**Тема 1.2. 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.

Багатовимірне шкалювання (БШ; [англ.](https://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D1%96%D0%B9%D1%81%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0) Multidimensional scaling; (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](https://uk.wikipedia.org/w/index.php?title=%D0%A2%D0%B0%D0%BA%D1%81%D0%BE%D0%BD%D1%96%D0%BC%D1%96%D1%8F&action=edit&redlink=1) 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]](https://uk.wikipedia.org/wiki/%D0%91%D0%B0%D0%B3%D0%B0%D1%82%D0%BE%D0%B2%D0%B8%D0%BC%D1%96%D1%80%D0%BD%D0%B5_%D1%88%D0%BA%D0%B0%D0%BB%D1%8E%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F#cite_note-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.

**Тема 1.3. 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.

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.

**Тема 1.4. 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. 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. ХХ 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.

**Тема 2.1. 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.

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.

**Тема 2.2. 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.

Координуючий шлях. Його ефект. Обчислення ковариаций (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).

**Тема 2.3. 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.

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).

**Тема 2.4. 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.

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.

**Тема 2.5. 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.

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 2.6. 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. 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".

Агломератівні 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 2.7. 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. 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 2.8. Поняття інтерпретації вихідних даних і основні методологічні принципи використання методів аналізу даних в соціології**

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.