**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ**

**«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Кафедра             соціології і публічного управління

(назва)

«**ЗАТВЕРДЖУЮ**»

Завідувач кафедри

В. М. Мороз

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(ініціали та прізвище) (підпис)

«31» серпня 2022 року

**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)

галузьзнань 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 (обов’язкова/вибіркова) / професійна підготовка (обов’язкова/вибіркова))

форма навчання денна

(денна/заочна)

Харків – 2022 рік

**Тема 1. 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. Агломератівні 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.** **Дані. Метадані (2 год.)**

1. Створення даних (Data Generation/Data Capture)

2. Обслуговування даних (Data Maintenance)

3. Синтез даних (Data Synthesis)

4. Використання даних (Data Usage)

5. Публікація даних (Data Publication)

6. Архівація даних (Data Archival)

7. Знищення даних (Data Purging)

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

**Базова література**

|  |  |
| --- | --- |
| 1 | Горбачик А.П., Сальнікова С.А. Аналіз даних соціологічних дослідженьзасобами SPSS: Навч. посіб.- Луцьк, 2008. – 164 с. |
| 2 | 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 |
| 3 | Паніотто В.І., Максименко В. С., Харченко Н.М. Статистичний аналізсоціологічних даних. - Київ, 2004. – 270 с. |
| 4 | 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**

|  |  |
| --- | --- |
| 5 | Лупан I.В., Авраменко О.В., Акбаш К.С. Комп’ютерні статистичні пакети: навчально-методичний посібник. - 2-е вид. - Кіровоград: ’КОД”. 2015. - 230 с. - http://dspace.cuspu.edu.ua/jspui/bitstream/123456789 |
| 6 | Making Sense of Multivariate Data Analysis//https://us.sagepub.com/en-us/nam/book/making-sense-multivariate-data-analysis |
| 7 | 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 |
| 8 | Інтелектуальний аналіз даних: практикум/ М.Т. Фісун, І.О. Кравець, П.П. Казмірчук.- Л.: Новий Світ-2000, 2020.- 162с. |
| 9 | Гладун А.Я., Рогушина Ю. В. Data Mining: пошук знань в даних. Київ. ТОВ «ВД «АДЕФ- Україна», 2016. — 452 с.. |

**ІНФОРМАЦІЙНІ РЕСУРСИ В ІНТЕРНЕТІ**

1. ukrstat.gov.ua (СтатистичніматеріалиДержкомстатуУкраїни)
2. http://i-soc.com.ua/institute/el\_library.php (Електронна бібліотека Інституту соціології НАН України)
3. <http://sau.in.ua/theme/simya/> (Соціологічна асоціація України, дослідження, сім’я)
4. <https://genderindetail.org.ua/> (Гендер в деталях)
5. <https://idss.org.ua/public> (Інститут демографії та соціальних досліджень імені М.В.Птухи Національної академії наук України, публікації)