***Тема* 4. Опрацювання та аналіз даних від збору до отримання цінної інформації (на прикладі реалізації публічних консультацій в інформаційній діяльності)**

***Мета:*** набути навички використання типів шкал та змінних, одновимірного аналізу та «крос-табуляції», вміти користуватися перевагами середнього значення, міри та медіани як мір центральної тенденції, структурувати результати оброблення даних у фінальному звіті та ефективно їх оприлюднювати.

*План*

1. Стратегія збору даних.

2. Опрацювання цінної інформації.

3. Структура для фінального звіту. Оприлюднення статистичних даних завдяки ефективним комунікаційним стратегіям.

***Основні поняття:*** стратегія збору даних, опрацювання інформації, оприлюднення статистичних даних, комунікаційні стратегії фінального звітування

**📚 Методичні рекомендації та поради**

У процесі підготовки самостійної роботи студентам необхідно взяти до уваги таку інформацію.

У посібнику від «Українського центру суспільних даних» [1] зазначається, що статистичні методи можуть бути використані для отримання відповідей на кшталт:

1. Якого типу дані і в якій кількості потрібно зібрати?
2. Як варто організувати та узагальнити дані?
3. Як ми можемо проаналізувати дані та дійти висновків?
4. Як ми можемо оцінити силу висновків та оцінити їхню непевність?

Стратегії збору даних ми вже частково розглядали в попередніх розділах, другому та третьому, тож зупинимося переважно на узагальненні та аналізі даних. Наступні тези щодо видів змінних, шкал для статистичного аналізу репрезентовано на основі матеріалів **«Відкритого посібника з відкритих даних для громадських організацій, журналістів, і не тільки…», упорядкованого Д. Кубаєм, А. Горбалем** на платформі «Українського центру суспільних даних» за посиланням: https://socialdata.org.ua/manual/ [1].

**Види шкал та змінних.**

Для цілей статистичного аналізу варто дефініціювати основні типи змінних (та типів шкал), аби зрозуміти, які математичні операції можна робити з різними змінними для статистичного аналізу.

**Змінна** – будь-яка характеристика об’єкта, що вимірюється чи досліджується, може набувати різних значень для різних об’єктів. Виділяють такі типи шкал:

**Номінальна шкала –** категорійна змінна, значення якої не можуть бути впорядковані – наприклад ім’я людини («Якби вибори Президента України відбулися у найближчу неділю, за кого б Ви проголосували?») або стать.Єдина операція, яку можна застосовувати до змінних з номінальною шкалою – це перелік, підрахунок частоти появи того чи іншого значення в масиві.

**Порядкова шкала** – змінна, значення якої за означенням впорядковані (наприклад «слабо», «помірно», «сильно»). Прикладом порядкової змінної в соціологічному опитування буде питання типу: «Наскільки ви довіряєте самому собі?» – з варіантами відповіді «Зовсім не довіряю», «Скоріше не довіряю, ніж довіряю», «Важко сказати довіряю чи ні», «Скоріше довіряю ніж не довіряю», «Цілком довіряю». Або ж значення в полі «Освіта» (неповна середня, середня, середня спеціальна, незакінчена вища, вища, вчений ступінь).

Значення такої змінної можна порівнювати на тотожність, на «більше-менше», але при цьому не можна сказати, **наскільки** більше чи менше.

**Метричні шкали:**

**Інтервальна – з**начення змінної впорядковані як в порядковій шкалі, але при цьому відмінності між значеннями є змістовними – наприклад, температура за Цельсієм, часові інтервали, тощо.

Допустимі операції: додавання та віднімання, але не множення чи ділення.

**Шкали відношень**

**Шкала вимірювань кількісної властивості** – змінні мають всі властивості, що й інтервальні, але також мають «природню» нульову точку «відліку» — наприклад вік, вага, температура за Кельвіном.

Додавання, віднімання, множення, ділення є *змістовними операціями*.

Також змінні розділяють на дискретні та неперервні. Так, кількість дітей у якійсь родині, кількість автомобільних аварій на певній території, або кількість людей, що прочитали цей текст за останній місяць – це результат простого підрахунку випадків настання певного явища, і тому є **значеннями дискретних змінних**. Дискретною є змінна, значеннями якої можуть бути лише цілі числа.

А **неперервні змінні** — такі як довжина, вага, температура, вік – можуть бути виміряні достатньо точно і не мають мінімальної неподільної одиниці виміру. Вага може бути виміряна з точністю до грама, а може й до мікрограма, а вік – з точністю до днів, годин, хвилин, секунд.

У невеликих масивах даних можна знайти мінімальне та максимальне значення, побачити, наскільки повторюються значення або наскільки багато значень є близькими один до одного. Для цього досить відсортувати масив даних в електронній таблиці за певною колонкою (змінною), тобто впорядкувати масив за певною ознакою. Але коли масиви даних надто великі, потрібно застосовувати інші методи для стиснення інформації, отримання розуміння про характеристики масиву.

**Описові статистики** єнайпростішими способами статистичного аналізу, які на найпростішому рівні кількісно описують або підсумовують масив даних.

**Одновимірний розподіл** – дозволяє підсумувати частоту, з якою різні значення певної змінної спостерігаються в наборі даних. Наприклад, ми маємо набір даних, у якому представлено 26 дітей віком від 1 до 6 років. Тоді **частотний розподіл змінної «вік» може матити в таблиці такий вигляд**:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Вік | 1 | 2 | 3 | 4 | 5 | 6 |
| Частота | 5 | 3 | 7 | 5 | 4 | 2 |

Звісно, коли змінні інтервальні, і значень їх дуже багато, доцільно розбити на певні інтервали, об’єднати значення змінних за певними категоріями:

|  |  |  |  |
| --- | --- | --- | --- |
| Вікова група | 1–2 | 3–4 | 5–6 |
| Частота | 5 | 3 | 7 |

Можна також побачити частоту того чи іншого значення у вигляді частки чи відсотка від усіх спостережень. У таблиці нижче представлено **частотний розподіл та відносний частотний розподіл.**

|  |  |  |
| --- | --- | --- |
| Оцінка студентів | Частота | Відносна частота |
| 0 – 25 | 2 | 0,13 |
| 26 – 50 | 5 | 0.31 |
| 51 – 75 | 4 | 0.25 |
| 76 – 100 | 5 | 0.31 |
| Разом | 16 | 1,00 |

Як видно з таблиці, підсумок відносної частоти в кожній із категорій дорівнює 1 – або 100% – оскільки включає в себе всі спостереження.

Одновимірний розподіл може бути реалізований щодо якісних та кількісних змінних, дискретних та неперервних, для відкритих та закритих категорій (так, **«відкритою» категорією** буде «люди старші 60 років», або «дохід нижче 3000 грн»). Ми можемо класифікувати дані **за кількісними характеристиками** – такими як вікові групи, зарплата, рівень доходу тощо; можемо також класифікувати **за якісними характеристиками** – такими як стать, професія, національність, тощо.

Приклад: **частотний розподіл кількісних дискретних даних з відкритою категорією**

| **Рівень доходу, грн** | **Частота** | **Відносна частота** |
| --- | --- | --- |
| 0 – 5000 | 15 | 0,30 |
| 5000 – 10000 | 25 | 0,50 |
| 10000 – | 10 | 0,20 |
| Разом | 50 | 1,00 |

Приклад: **частотний розподіл даних номінальної шкали виміру**

| **Стать** | **Частота** | **Відносна частота** |
| --- | --- | --- |
| Чоловіки | 20 | 0,40 |
| Жінки | 30 | 0,60 |
| Разом | 50 | 1,00 |

**Двовимірний розподіл або крос-табуляція –** ще один спосіб узагальнення даних в таблиці. Це таблиця, що представляє дані, зіставляючи дві окремі характеристики (змінні) – по колонках і рядках. Так, в колонках – різні значення (чи групи значень, категорії) однієї змінної, а в рядках – значення іншої змінної.

Приклад: таблиця з даними Держстату

| **Область** | **2012** | **2013** | **2014** |
| --- | --- | --- | --- |
| Вінницька | 35441 | 37323 | 39184 |
| Волинська | 19546 | 20609 | 21971 |
| Дніпропетровська | 95349 | 99995 | 109545 |
| Донецька | 128767 | 135362 | 114135 |

Важливо розуміти, що з точки зору грамотного і охайного структурування даних – до початку аналізу – у наборі даних (у таблиці) всі змінні мають бути в колонках, всі спостереження в рядках, а на їх перетині – значення змінних для спостережень.

Але для аналізу даних можна порівнювати одні змінні з іншими, **робити крос-табуляцію**, **будувати двовимірні розподіли** – наприклад, для того, щоб на найпростішому рівні виявити залежності між змінними.

Приклад: два одновимірних розподіли наведених вище – за доходом і статтю.

| **Стать** | **Дохід** | **Разом у рядках** |
| --- | --- | --- |
|  | 0 – 5000 | 5000 – 10000 |
| Чоловіки | 7 | 9 |
| Жінки | 8 | 16 |
| Разом у колонках | 15 | 25 |

**Будуючи таблицю розподілу для набору даних, що містить велику кількість спостережень, варто дотримуватися таких основних правил** (за Д. Кубаєм, А. Горбалем):

* знайти найменше та найбільше значення змінної;
* визначити «ширину» категорій, за якими будуть згруповані значення змінної;
* включити всі можливі значення змінної.

Визначаючись із шириною інтервалу, потрібно знайти компроміс: щоб інтервали були достатньо вузькими (щоб всі або більшість спостережень не потрапили в один інтервал), і при цьому достатньо широкими (щоб не вийшло так, що буде по одному спостереженню в інтервалі). Також важливо, щоб інтервали були взаємно виключними.

Міри центральної тенденції – **середнє**, **моду** та **медіану** – ми вже розглядали у змістовому модулі 2, а саме їхню механіку, а ось правила, переваги та недоліки варто розглянути на основі тез «Відкритого посібника з відкритих даних для громадських організацій, журналістів, і не тільки…» [1].

Ці показники демонструють загальні характеристики розподілу даних за певною змінною, дозволяють виявити одне значення, що описує весь розподіл; це окремі значення, що представляють весь набір даних, типові для всіх значень у групі.

**Міри центральної тенденції потрібні з таких міркувань:**

1. Щоб отримати загальну картину розподілу. Ми не можемо запам’ятати кожен факт, що стосується сфери дослідження.
2. Щоб отримати чітку картину щодо досліджуваної сфери для розуміння та отримання потрібних висновків.
3. Щоб отримати чіткий опис групи в цілому та мати змогу порівнювати дві або більше груп у термінах типової «поведінки».

Найвідомішою мірою центральної тенденції — і найбільш вживаною в повсякденному побуті – є **середнє**, або ж просте середнє, або ж арифметичне середнє (arithmetic mean) – просто середнє значення ряду даних.

Для його обчислення досить скласти разом всі значення в розподілі і поділити на кількість спостережень. В Екселі чи Google Spreadsheets для цього є функція MEAN. Є різні математичні способи підрахунки середнього, але в усіх сучасних електронних таблицях та спеціальних програмних пакетах для роботи з даними і статистикою є ця функція.

**Є певні загальні правила для використання середнього, зокрема:**

1. Середнє – це «центр тяжіння» розподілу, і кожне значення дає внесок у визначення середнього значення, коли поширення значень є симетричними довкола центральної точки.
2. Середнє значення більш стабільне, ніж медіана чи мода. Тому, коли потрібно знайти найбільш стабільну міру центральної тенденції, використовують середнє.

**Переваги середнього:**

1. Середнє визначене дуже жорстко, тому не виникає питань чи нерозуміння щодо його значення та суті.
2. Це найбільш поширена міра центральної тенденції, оскільки її легко зрозуміти.
3. Середнє легко підрахувати.
4. Враховує всі значення розподілу.

**Обмеження чи недоліки середнього:**

1. На значення середнього впливають екстремальні значення (відомий іронічний жарт про «середню температуру по лікарні»).
2. Часом середнім є значення, що не присутнє в розподілі.
3. Часом результатом можуть бути абсурдні значення. Наприклад, маємо 41, 44, та 42 учнів у 5а, 5б та 5в класах певної школи. Виходить, що середня кількість учнів у 5 класах школи – 42,3 (3). А так не буває.

**Медіану** можна визначити як точку на ряді розподілу (впорядкований набір значень змінної для різних спостережень – наприклад від найменшого до найбільшого значення) – до цієї точки розташовано половина всіх значень, і після цієї точки теж половина значень. Тобто медіана – це значення, що ділить впорядкований ряд навпіл. Якщо кількість значень непарна, то береться одне зі значень – те, що стоїть у розподілі рівно по центру.

Коли значень парна кількість, то беруть два центральні значення і знаходять їхнє середнє.

**Для чого використовують медіану?**

1. Коли потрібно знайти точну середню точку, точку на «півдорозі» від найменшого значення до найбільшого.
2. Коли екстремальні значення впливають на середнє, то медіана є найкращою мірою центральної тенденції.
3. Медіану використовують, коли потрібно, щоб певні значення впливали на центральну тенденцію, але все, що про них відомо, що вони «нижче» або «вище» медіани.

**Переваги медіани:**

1. Легко вирахувати та зрозуміти.
2. Для підрахунку медіани не потрібні всі значення в розподілі.
3. Екстремальні значення розподілу не впливають на медіану.
4. Її можна визначити і для «відкритих» категорій / класів інтервалів.

**Обмеження медіани:**

1. Вона не так жорстко визначена як середнє, оскільки її значення не так вираховується, як знаходиться (серед значень в розподілі).
2. Не враховує всі спостереження (значення для всіх спостережень).
3. З медіаною потім не можна робити алгебраїчні перетворення так, як із середнім.
4. Потребує впорядкування значень або класів інтервалів у висхідному чи спадному порядку.
5. Часом медіаною може бути значення, не присутнє у самому розподілі.

Третя міра центральної тенденції – це **мода** – значення, що найчастіше зустрічається в розподілі. Як правило, вона представляє найбільш типове значення. На моду ніколи не впливають екстремальні значення в розподілі, а впливають – екстремальні частоти значень, наскільки часто те чи інше значення змінної зустрічається в розподілі.

**Мода використовується:**

1. Коли нам треба швидка і приблизна міра центральної тенденції.
2. Коли потрібна міра центральної тенденції, що має бути типовим значенням.

**Переваги моди:**

1. Мода показує найбільш поширене значення в розподілі.
2. На моду не впливають екстремальні значення – так як на середнє.
3. Моду можна визначити для відкритих інтервалів / категорій.
4. Допомагає аналізувати якісні дані.
5. Моду можна виявити просто побудувавши графік розподілу чи стовпчасту діаграму.

**Обмеження:**

1. Не включає до визначення / розрахунку всі спостереження розподілу, а лише концентрацію частот.
2. Подальші алгебраїчні перетворення неможливі – на відміну від середнього.
3. Буває важко визначити моду у випадку багатомодального чи бімодального розподілу

Розподіл може мати більше двох популярних значень, але якщо має більше ніж трьох мод, опис такого розподілу в термінах найбільш частих значень може втрачати будь-який сенс.

Д. Кубай, А. Горбаль як упорядники посібника наголошують на активному використанні графічного відображення, візуалізації даних, без якої статистичний аналіз є неможливим фактично.

Таким чином, графічне відображення стає важливим інструментом у підготовці детального звіту за підсумками отриманої інформації та презентації висновків.

Наприклад, після завершення публічних консультацій, які ми обрали як «робочий дискурс», у рамках якого ми проводимо аналіз даних, потрібно підготувати детальний звіт, який міститиме результати, що доцільно надати тим, хто ухвалює рішення, а також поширити на широкий загал. Такий документ-звіт є дуже корисним, оскільки:

* «він чітко описує те, як внесок громадян використають для розробки політики;
* він додає прозорості та легітимності публічним консультаціям, що може переконати інших людей долучатися до такої діяльності в майбутньому;
* він наголошує на тому, яку кропітку роботу виконали під час проведення консультацій;
* у ньому документують процес публічних консультацій, пропонують можливості для поступового вдосконалення практик, методів та засобів;
* і найважливіше — він слугує інструментом розробки політики» [2].

**Орієнтовна структура для фінального звіту** за [2]:

1) *вступ* (коротке пояснення, про що йдеться у звіті та чому уряд, установа взялися проводити консультації. Ця частина дає відповідь на запитання: «Яку проблему хотів вирішити уряд/організація/установа?»;

2) *контекст* (підсумовують довідкову інформацію, яку надавали громадянам, щоб забезпечити їхнє розуміння теми консультацій та сприяти їхній змістовній участі в обговореннях);

3) *процес проведення консультацій* (перераховують заходи, проведені протягом усього процесу консультацій, описують консультаційні події, комунікаційні стратегії та перераховують залучених партнерів, пояснюють методологію, яку використали для збору й аналізу даних);

4) *учасники* (детальна інформація про те, хто брав участь у консультаціях і наскільки добре вони представляли вашу цільову авдиторію);

5) «*що ми почули*» (найважливіший у звіті й ключовий при формуванні політики фрагмент, де узагальнюють коментарі та дискусії учасників, виокремлюють важливі висновки та групують провідні теми консультацій, ґрунтується на отриманих даних);

6) *рекомендації щодо розробки політики та засвоєні уроки* (різні рекомендації за результатами публічних консультацій; інформування щодо наступних кроків процесу розробки політики, якщо вони вже відомі; зазнаються ті уроки, що засвоїли під час публічних консультацій для удосконалення процесу в майбутньому);

7) *додатки* (всі документи, що мають відношення до проведених консультацій, зокрема офіційні документи, копії онлайн-опитувальників, приклади оголошень, розроблених для контактування з цільовою авдиторією тощо) [2].

«Щоб визначити, хто має бути присутній на цій зустрічі, варто проаналізувати, хто має повноваження затверджувати рішення та коли, а також звернути увагу на ключові моменти ухвалення рішень, зокрема зустрічі з обговорення політики, процеси планування бюджету, експертів, що проводили збір даних, експертів тематичної проблематики, представників комунікаційного відділу міністерства, організації тощотощо» [2].

**Звіт** варто опублікувати за підсумками проведених консультацій **на урядовому вебсайті, на офіційному сайті установи, організації**, де будуть розміщені звіт та всі відповідні консультаційні документи. Таку сторінку можна в подальшому оновлювати на кожному етапі процесу розробки політики, як-от подання законопроєкту до парламенту, його затвердження або імплементація. Пошук цієї інформації має бути легким для громадськості та всіх, хто брав участь у консультаціях,

Утім, буде недостатнім просто оприлюднити звіт за результатами консультацій у мережі. Треба застосувати ефективні комунікаційні стратегії й переконатися, що стейкхолдери та широкий загал дізнаються про результати консультацій. У посібнику «Публічні консультації: крок за кроком» [2] пропонують такі варіанти:

* організувати **пресконференцію** для презентації підсумків консультацій;

зняти **відеоролик** та/або розробити **брошуру,** в яких узагальнені ключові моменти проведених консультацій;

* опублікувати **звіт** у соціальних мережах та на будь-яких каналах, що виявилися успішними для залучення стейкхолдерів на попередніх етапах **консультацій**;
* роздрукувати **примірники звіту** та розіслати їх відповідним групам і обʼєднанням.

**Інформаційно-комунікаційна робота** не закінчується оприлюдненням та поширенням фінального звіту, оскільки у стейкхолдерів можуть виникнути подальші запитання або зауваження. Публічні консультації спрямовані на підтримку діалогу, тому стейкхолдерам іноді пропонують залишатися залученими за допомогою різних механізмів комунікації, як-от:

* інформаційні бюлетені;
* соціальні медіа та застосунки для обміну повідомленнями;
* групи для обговорень.

Постійна комунікація сприяє участі в майбутніх консультаціях, а залучені стейкхолдери можуть стати союзниками у просуванні та реалізації нової політики.

Приклади системи збору та аналізу даних для прийняття стратегічних рішень у різних царинах презентовано в додаткових джерелах використаної літератури до теми.

**❓***Питання для самоконтролю:*

1. Які типи шкал виділяють у статистичній практиці?

2. Розтлумачте поняття «одновимірний аналіз» та «крос-табуляція».

3. Яких основних правил варто дотримуватися при побудові таблиці розподілу для набору даних, що містить велику кількість спостережень?

4. Назвіть переваги, обмеження та недоліки середнього значення, міри та медіани як мір центральної тенденції.

5. Окресліть складники орієнтовної структури для фінального звіту публічних консультацій.

***Рекомендації до виконання завдань***: насамперед варто переглянути визначену до плану заняття *основну літературу* та *додаткову*.

**✍** *Завдання для самостійної роботи студента:*

1. Опрацювати онлайн-посібник «Публічні консультації: крок за кроком» https://edera.gitbook.io/publ-chn-konsultacz-krok-za-krokom/: розділ «Проведення публічних консультацій. Крок за кроком: Етап 6, 7», заповнивши шаблони, які там пропонуються. Сформувати звіт за проаналізованими даними. Презентувати проєкт.

2. Пройти онлайн-курс «Публічні консультації» на платформі Edera за посиланням: https://courses.ed-era.com/courses/course-v1:NACS\_NDI\_EDERA+PC101+2019/about.

**🕮 *Література:***

***Основна:***

1. Кубай Д., Горбаль А. Відкритий посібник з відкритих даних для громадських організацій, журналістів, і не тільки… Київ : Український центр суспільних даних, 2016. URL: https://socialdata.org.ua/manual/ (дата звернення: 08.04.2023).

2. Публічні консультації: крок за кроком : посібник. *Edera.* URL: https://edera.gitbook.io/publ-chnkonsultacz-krok-za-krokom/provedennya-publichnikh-konsultacii.-krok-za-krokom/etap-5.-proanalizuite-otrimani-dani (дата звернення: 08.04.2023).

***Додаткова:***

3. Боброва А., Вербицький І. Дані та статистика про міста і громади: основні тези дискусії. URL: https://mistosite.org.ua/articles/dani-ta-statystyka-pro-mista-i-hromady-osnovni-tezy-dyskusii (дата звернення: 08.04.2023).

4. Вдосконалення розвитку статистики в Україні. *Методи і стандарти розробки запитальників і анкет.* 13 липня 2021 року. URL: https://www.oecd.org/eurasia/countries/ukraine/Draft\_Proceedings\_Webinar2\_StatisticsUKR.pdf (дата звернення: 08.04.2023).

5. Семінар щодо вдосконалення системи збору та аналізу судових даних для прийняття стратегічних рішень. URL: https://www.coe.int/uk/web/kyiv/-/workshop-on-improving-the-collection-and-analysis-of-judicial-data-for-policy-making-purposes (дата звернення: 08.04.2023).

6. Система збору даних про злочини на ґрунті ненависті, мову ворожнечі та дискримінацію в Україні: рекомендації щодо вдосконалення та застосування спільного підходу та ситуаційний аналіз. вересень 2020 р. *Проєкт «Посилення доступу до правосуддя через несудові механізми правового захисту для жертв дискримінації, злочинів на ґрунті ненависті та мови ворожнечі в країнах Східного партнерства»* URL: https://rm.coe.int/final-data-collection-report-ukraine-ua/16809fac71 (дата звернення: 08.04.2023).