***Тема* 2. Методологічні засади статистики. Статистичне спостереження**

***Мета*:** визначати типи даних, користуватися основними категоріями статистики, статистичного спостереження, формувати матрицю даних.

*План*

1. Процес аналізу даних. Типи даних. Матриця даних.

2. Види статистики. Основні категорії статистики.

3. Статистичне спостереження: форми, види, способи.

 **Основні поняття:** статистика, статистичне спостереження, матриця даних, типи даних, середнє значення, мода, медіана

**📚 Методичні рекомендації та поради**

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

 **Аналіз даних** – це процес роботи з даними з метою їх правильного розміщення, пояснення, презентабельності та пошуку висновку із цих даних для винайдення корисної інформації для прийняття раціональних рішень.

Відповідно, **основною метою аналізу даних** є інтерпретація, оцінка, організація даних, а також їх ефективна та ефектна презентація.

 **Процес аналізу даних** включає:

* Збір даних;
* Робота над якістю даних;
* Побудова моделі;
* Навчальна модель;
* Запуск моделі з повними даними.

Деякі поради щодо аналізу даних:

* Видаліть непотрібні дані перед аналізом.
* Не слід проводити аналіз на головній копії даних.

 За даними спеціалізованого порталу myservername.com, варто розрізняти аналіз даних, видобуток даних та моделювання даних.

 «**Аналіз даних** проводиться з метою пошуку відповідей на конкретні питання. Методи аналізу даних подібні до бізнес-аналітики та бізнес-аналітики.

 **Видобуток даних** – це пошук різних моделей даних. Для цього до даних застосовуються різні математичні та обчислювальні алгоритми й генеруються нові дані.

 **Моделювання даних** – це те, як компанії впорядковують дані або керують ними. Тут до даних застосовуються різні методології та методи. Для моделювання даних необхідний аналіз даних» [3].

Порівняння найкращих інструментів аналізу даних репрезентовано у Додатку А.

Як засвідчує А. Корнілова, для аналізу даних необхідно три основні компоненти:

* знання предметної області (це дозволяє розуміти, які проблеми потребують першочергового вирішення);
* знання математики та статистики (дозволяють формалізувати рішення, перевести його в алгоритм та оцінити, яка ймовірність отримати результат);
* вміння програмувати (дає можливість застосовувати величезні обчислювальні потужності) [1].



Рис. 6. Компоненти, необхідні для аналізу даних. Джерело: [1]

**Процес аналізу даних складається з трьох етапів.**

«Спочатку дані потрібно підготувати, тобто зібрати, очистити та відібрати ті, які потрібні для моделі. Цей процес займає близько 90% часу. Далі ми будуємо модель та валідуємо її результати. Останній етап – це презентація результатів. Тут ми демонструємо на яке питання ми шукали відповідь, які дані використовували та що отримали в результаті. Для того щоб це зробити максимально ефективно треба витрати ще 90% часу» [1].



Рис. 7. Етапи процесу аналізу даних. Джерело: [1]

Існує **два методи аналізу даних**: якісний аналіз та кількісний аналіз.

* Якісний аналіз: якісний аналіз проводиться за допомогою інтерв’ю та спостережень.
* Кількісний аналіз: кількісний аналіз проводиться за допомогою опитувань та експериментів.

Відповідно, поняття «**статистика**» (від лат. status – стан речей) означає кількісний облік масових, насамперед соціально-економічних явищ і процесів.

Як окрема галузь науки статистика виникла з повсякденних практичних потреб людей, оскільки для державного управління потрібна інформація, наприклад, про наявність і склад земель, чисельність і склад населення, стан торгівлі, доходи і витрати населення, рівень і динаміку цін на товари та послуги, розвиток матеріального виробництва тощо; тобто як наука статистика об’єднує принципи та методи роботи з масовими числовими даними, тобто кількісними характеристиками зазначених явищ і процесів.

*Об’єкт її вивчення* – соціальні, економічні, політичні та культурні явища і процеси суспільного життя.

*Предмет* – розміри і кількісні співвідношення між масовими суспільними величинами, закономірність їх формування, розвитку і взаємозв’язку. Тобто, по-перше, статистика вивчає кількісний бік суспільних явищ, а по-друге, вона вивчає не поодинокі, а масові явища.

Статистика в сучасних умовах забезпечує державні органи всіх рівнів інформаційно аналітичними матеріалами, на основі яких розробляється податкова та цінова політика, приймаються засоби зі стимулювання або стримання розвитку ринку та окремих його сегментів, забезпечується соціальний захист населення і т. д. Комерційні підрозділи та спеціалізовані маркетингові фірми, поряд з обов’язковою звітністю, формують панелі споживачів, займаються опитуванням та анкетуванням, тобто всіма засобами вивчають цільову авдиторію та конкурентів за всіма параметрами.

Статистичні органи на комерційній основі надають фірмам, установам, організаціям статистичну інформацію – ряд відомостей, які утворюють, наприклад, маркетингове середовище (демографічні дані споживачів, індекси цін, індекси вартості життя тощо); формують прогнози, оцінки

**Основні категорії статистики**, за допомогою яких відображаються найбільш узагальнювальні та суттєві властивості, ознаки, зв’язки та відношення предметів та явищ (у базових наукових розвідках їх вивчали В.Л. Бек, А.Г. Гончарук, В.К. Горькавий):

1. **Статистична** [**сукупність**](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=2440&displayformat=dictionary) – це певна множина елементів, поєднаних умовами існування і розвитку. Так, наприклад, статистичною є [сукупність](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=2440&displayformat=dictionary) домогосподарств, [сукупність](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=2440&displayformat=dictionary) родин, [сукупність](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=2440&displayformat=dictionary) підприємств, фірм, об’єднань тощо. Сукупності можуть бути *однорідними* (якщо одна або декілька суттєвих ознак її об’єктів є узагальнювальними для всіх [одиниць](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=2454&displayformat=dictionary)) та *різнорідними* (до сукупності входять явища різного типу).

* **Одиниця сукупності –** окремі об’єкти (елементи, явища тощо), які становлять статистичну [сукупність](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=2440&displayformat=dictionary), характеризуються індивідуальними ознаками.

Так, наприклад, ознаками торгівельного підприємства є розміри основних та оборотних засобів, обсяг та структура реалізованої продукції тощо. Ознаками продукції є якість, термін зберігання, вартість і т. п.

Одні ознаки виражають числами, а інші – вербально. Їх називають відповідно якісними і атрибутивними (описовими).

Особливістю **статистичного дослідження** є те, що в ньому вивчаються тільки [варіативні ознаки](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=3383&displayformat=dictionary), тобто які мають в межах сукупності різні значення. Відмінність, коливання значення ознаки називається **варіацією.** Наприклад, ознаки підприємства: спеціалізація, форма власності, [рентабельність](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=3461&displayformat=dictionary) виробництва тощо.

Якщо ж зміна явища відбувається в різні періоди часу, зберігаючи рівномірний характер, то говорять вже не про варіацію ознаки, а про її **динаміку**. **Статистичні показники** можуть бути *об’ємними* (чисельність населення, трудових ресурсів) та *розрахунковими* (середні величини), *плановими*, *звітними*, *прогностичними*.

За формою і змістом статистичні показники поділяються на *абсолютні* та *відносні*. Абсолютні показники характеризують розмір суспільних явищ, наприклад, розмір посівних площ, обсяг виробленої та реалізованої продукції, суму прибутку, вартість основних фондів, чисельність робітників, фонд заробітної плати підприємства тощо. Вони відповідають на запитання «скільки?» і завжди відображаються іменованими числами (га, ц, т, грн). Відносні показники відображають кількісні співвідношення між абсолютними показниками (напр. у відсотках, коефіцієнтах).

При розрахунку відносного показника абсолютний показник, що знаходиться в чисельнику одержуваного відносного, називається *поточним або порівнюваним.* Показник же, з яким проводиться порівняння і який знаходиться в знаменнику, називається *підставою або базою порівняння*. Таким чином, розраховувана відносна величина показує, у скільки разів порівнюваний абсолютний показник більше базисного, або яку становить від нього частку, або скільки одиниць першого припадає на 1, 100, 1000 і т. д. одиниць другої.

Так, собівартість виробництва 1 т продукції, грн (відносний показник) отримують внаслідок ділення собівартості виробництва всієї продукції, грн на кількість виробленої продукції, т.

Відносний показник виконання плану – це відношення фактично досягнутого рівня до планового завдання. У нижче наведеному прикладі показник відображає фактичний обсяг виробництва або реалізації у відсотках або коефіцієнтах порівняно з плановим рівнем.



**Відносні показники мають велике аналітичне значення**, їх обчислюють, щоб мати кількісну характеристику різноманітних сторін суспільного життя. За їхньою допомогою виражають ступінь виконання планів, ефективність та інтенсивність суспільного виробництва, продуктивність праці, ступінь задоволення матеріальних і культурних потреб людей, структуру і динаміку виробництва тощо. Завдяки ним можна порівнювати однойменні і різнойменні величини.

**Система статистичних показників** – це [сукупність](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=2440&displayformat=dictionary) статистичних показників, що відображає взаємозв’язків, які об’єктивно існують між явищами.

Система статистичних показників охоплює всі напрями життя суспільства на різних рівнях: країни, регіонів – макрорівень; підприємств, фірм, об’єднань, родин, домогосподарств і т. п. – мікрорівень.

Системи статистичних показників мають такі особливості:

* вони мають історичний характер – змінюються умови життя населення, суспільства, змінюються і системи статистичних показників;
* методологія розрахунку статистичних показників безперервно удосконалюється.

Існують **три групи методів статистичного дослідження**:

[**Статистичне спостереження**](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=3469&displayformat=dictionary) – дає інформаційну базу для статистичних узагальнень та характеристики об’єктивних [закономірностей](https://cdn.snau.edu.ua/moodle/mod/glossary/showentry.php?eid=2447&displayformat=dictionary) ( тобто на цьому етапі здійснюється збирання первинного статистичного матеріалу). Це перший етап будь-якого статистичного дослідження.

**Метод групувань** – другий етап статистичного дослідження. На цій стадії зібрані факти класифікуються і систематизуються.

**Метод узагальнювальних показників** дає змогу надати характеристику досліджуваних явищ та процесів за допомогою статистичних величин – абсолютних, відносних, середніх. На цьому етапі статистичного дослідження виявляються взаємозв’язки та масштаби явищ, визначаються закономірності їх розвитку, даються оцінки на майбутнє.

**Закономірність** – це повторюваність, послідовність та порядок у розвитку соціальних явищ. Вона може проявлятися по-різному. Розрізняють два види закономірностей: *динамічну* та *статистичну.*

У природних явищах спостерігаються закономірності, які звуться динамічними, тобто певні ознаки того чи іншого явища спостерігаються у кожному окремому його прояві, зокрема, у фізичних, хімічних, математичних законах. Скажімо, площа кола змінюється на відповідну величину у зв’язку зі зміною його радіусу, закон Архімеда проявляється у кожному випадку занурення тіла у рідину тощо.

У суспільному житті в соціальних явищах такої закономірності не спостерігається, хоча вони в таких явищах теж існують. Але свій прояв закономірності соціальних явищ знаходять при масовому і послідовному спостереженні, і тому такі спостереження звуться *статистичними,* щоне виражають властивості кожного явища окремо, а тільки в масових процесах або в загальній системі.

**Статистична закономірність** – це така, яка виявляється лише в достатній кількості однорідних одиничних елементів, котрі й утворюють сукупність. Тобто кожний окремий елемент може не підтверджувати існування тієї чи іншої закономірності, тому що існування її в кожному елементі носить імовірний характер. Інакше кажучи, статистична закономірність властива лише сукупності одиниць, яка має назву статистичної сукупності, проявляється лише на підставі дослідження достатньо великої кількості одиниць спостереження. Наприклад, залежність злочинності від доходів родини, культурного рівня населення тощо.

Отже, як зауважує А. Корнілова, «статистика допомагає оцінити варіативність та зменшити невизначеність. Розрізняють описову та вивідну статистики.

* Описова – вивчає властивості спостережуваних даних.
* Вивідна статистика – виводимо припущення про властивості розподілу даних, з яких походять спостережувані дані.

За допомогою статистики можна дати відповідь на питання:

- чи є залежність між кількістю злочинів та фазою Місяця?

- яка ймовірність викликати Uber в Києві?

- побудувати довірчий інтервал часу, за який ви потрапляєте на роботу

- проводити опитування та трактувати їх результати» [1].



Рис. 7. Типи даних. Джерело: [1]

Матриця даних – «стартовий елемент для аналізу даних. Зазвичай йому передує етап збору, очищення та представлення у табличному вигляді. По рядках – респонденти, суб’єкти, учасники, спостереження По стовпцях – xарактеристики кожного запису (змінні). Також важливо звертати увагу на одиниці виміру, а також яким чином були зібрані ці дані. Ця таблиця включає в себе 6 рядків, однак зібрані дані мають майже 800 спостережень (спостереження зібрані з ресурсу і містять інформацію про квартири, які продаються). Для того, щоб описати вміст цієї таблиці в більш зрозумілій формі, використовують узагальнення та опис типових чи середніх значень. Для цього важливо знати тип даних» [1] (див. Рис. 8).



Рис. 8. Матриця даних. Джерело: [1]

Для узагальнення категоріальних даних використовують частотні таблиці. Наступна таблиця, презентована в рамках онлайн-курсу «Аналіз даних та статистичне виведення на мові R», містить кількість квартир, що продаються у кожному місті [1] (див. Рис. 9).



Рис. 9. Частотні таблиці. Джерело: [1]

**Центральна тенденція**. «Центральне або типове значення дозволяє зрозуміти основну характеристику даних.

**Середнє значення** підходить для узагальнення кількісних даних (як дискретних, так і неперервних). Формула обрахування проста: суму всіх чисел ділимо на їх кількість. Наприклад, якщо в нас є група з 5 учнів, оцінки яких 12, 3, 5, 10, 5. Сума їх оцінок дорівнює 35, а середнє значення 7. Однак із використанням середнього значення як опис центральної тенденції в даних є невелика проблема. Якщо є нетипово великі чи малі для даного набору значення – вони роблять великий внесок у значення середнього. Нехай у нас є певне невелике підприємство, яке має 5 працівників. Заробітні плати працівників в гривнях: 5000, 7000, 2000, 4000, 50 000. Середнє значення заробітної плати 13600 грн. Однак, якщо ми відкинемо екстремальне значення 50 000, то отримаємо, що середнє значення зменшилося до 4500» [1] (див. Рис. 10).

.

Рис. 10. Середнє значення. Джерело: [1]

**Медіана** – «це значення, яке ділить вибірку навпіл, тобто 50% є меншими за це значення, 50% більшими. Основна перевага використання медіани – менша чутливість до екстремальних значень. Для пошуку медіани дані треба розмістити в зростаючому порядку та поділити на дві частини. Якщо в нас парна кількість спостережень, то сусідні значення по краях сумуються та діляться на два. У випадку попереднього прикладу із заробіною платою: 2000, 4000, 5000, 7000, 50 000. Маємо, що по середині знаходиться значення 5000, то краще описує центральну тенденцію заробітної плати на підприємстві» [1] (див. Рис. 11).



Рис. 11. Середнє значення. Джерело: [1]

 **«Мода** використовується для визначення центральної тенденції категоріальних або кількісних дискретних даних. Мода – це значення, яке найчастіше трапляється. Наприклад, за інформацією Міністерства юстиції України у певний відрізок часу хлопчиків найчастіше називали Дмитром, Артемом, Максимом та Іваном, дівчаток – Анею, Анастасією, Софією та Дар’єю. (http://tyzhden.ua/News/137908). Ці імена були модою серед всіх імен» [1] (див. Рис. 12).



Рис. 12. Мода. Джерело: [1]

Власне, у статистичній практиці застосовують дві організаційні форми спостереження – **звітність** і **спеціально організовані статистичні спостереження** [2, с.19].

*Звітність* – це форма спостереження, при якій кожний суб’єкт діяльності подає свої дані в державні органи статистики та відомства у вигляді документів, звітів спеціально затвердженої форми.

*Спеціально організоване спостереження* – це форма спостереження, яка охоплює сфери життя та діяльності, що не відображуються звітністю. До числа таких спостережень належать: переписи, обліки, спеціальні обстеження, опитування.

*Перепис* – суцільне або вибіркове спостереження окремих масових явищ з метою вивчення їхнього розміру та складу на певну дату.

*Обліки* – суцільне спостереження окремих масових явищ згідно з певною тематикою, що виходить за межі звітності; можуть бути періодичними або одноразовими.

*Спеціальні обстеження* – несуцільні спостереження окремих масових явищ згідно з певною тематикою, що виходить за межі звітності; можуть бути періодичними та одноразовими.

*Опитування* – це, як правило, несуцільне спостереження за думками, мотивами, оцінками, що реєструються зі слів респондентів

**❓***Питання для самоконтролю:*

1. Назвіть етапи процесу аналізу даних.

2. Розкрийте основні категорії статистики та статистичного дослідження.

3. Які організаційні форми спостереження застосовують у статистичній практиці?

4. Які формули вирахування середнього значення, моди та медіани даних?

***Рекомендації до виконання завдань***: насамперед варто переглянути визначену до плану заняття *основну літературу* та *додаткову*. Важливо у процесі підготовки користуватисяматеріалами онлайн-курсу «*Аналіз даних та статистичне виведення на мові R» https://courses.prometheus.org.ua/courses/IRF/Stat101/2016\_T3/about*, а також звертатися до *категоріального апарату*.

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

1. Виконати завдання із визначення середнього значення, медіани, моди в заданій матриці даних (тема та дані обираються довільно).

2. Визначити актуальні тренди в конкретному регіоні в режимі реального часу. Скористайтеся інструментом «Трендові запити» для пошуку популярних тем та інфоприводів для інтернет-видань, тематичних контент-платформ та нішевих медіа: http://surl.li/gdawt.

3. Опрацювати статтю «Статистика відвідувань сайту»: https://wezom.com.ua/ua/blog/statistika-poseshchenij-sajta.

Скористатися одним із запропонованих інструментів.

Або проаналізувати статистику своєї сторінки в соціальній мережі (напр., за інструментарієм, описаним за посиланням: https://www.facebook.com/help/268680253165747?cms\_platform=www&helpref=platform\_switcher)

4. Апробувати ресурси з визначення індексів, рейтингів тощо, сформувати інформаційну довідку щодо знайдених статистичних показників, даних (у списку рекомендованої літератури [8–20].

**🕮 *Література:***

 ***Основна:***

1. Корнілова А. Prometheus. Аналіз даних та статистичне виведення на мові R : конспект лекцій. URL: http://surl.li/gdbhf (дата звернення: 08.04.2023).

2. Соціальна статистика : конспект лекцій для здобувачів вищої освіти першого (бакалаврського) рівня зі спеціальності 054 «Соціологія» / укладач А.В. Хмелюк. Кам’янське : ДДТУ, 2020. 155 с.

 ***Додаткова:***

3. 10 найкращих інструментів аналізу даних для ідеального управління даними [Список 2021]. URL: https://uk.myservername.com/10-best-data-analysis-tools (дата звернення: 08.04.2023).

4. Job Comparison – Data Scientist vs Data Engineer vs Statistician. Published On October 19, 2015 and Last Modified On June 24th, 2019. Analytics vidhya. URL: https://www.analyticsvidhya.com/blog/2015/10/job-comparison-data-scientist-data-engineer-statistician/ (дата звернення: 08.04.2023).

5. Nathan Yau. Flowing Data. URL: https://flowingdata.com/ (дата звернення: 08.04.2023).

6. 100 interesting data sets for statistics. URL: https://rs.io/100-interesting-data-sets-for-statistics/ (дата звернення: 08.04.2023).

7. R for Data Science. URL: https://r4ds.had.co.nz/index.html (дата звернення: 08.04.2023).

**Інформаційні ресурси:**

8. Державна служба статистики України. Статистика щодо економічної ситуації в Україні: ВВП, інфляція, безробіття, експорт / імпорт, реєстр підприємств, сільське господарство, будівництво, регіональна статистика та ін. URL: http://www.ukrstat.gov.ua/ (дата звернення: 08.04.2023).

9 Державний і місцеві бюджети: доходи, витрати, обслуговування , державного боргу, виконання бюджету. URL: http://www.treasury.gov.ua/ (дата звернення: 08.04.2023).

10. Євростат. Статистика з основних макропоказників в СС (ВВП, інфляція, безробіття та ін.). URL: http://ec.europa.eu/eurostat (дата звернення: 08.04.2023).

11. Єдиний веб-портал використання публічних коштів. URL: https://spending.gov.ua/new/ (дата звернення: 08.04.2023).

12. Індекс добробуту по країнах. URL: http://worldhappiness.report/ (дата звернення: 08.04.2023).

13. Індекс політичного розвитку в державах світу. URL: https://ec.europa.eu/ (дата звернення: 08.04.2023).

14. Індекс розвитку свободи слова. URL: https://rsf.org (дата звернення: 08.04.2023).

15. Індекс свободи ведення бізнесу, умови підприємництва в країнах. URL: http://www.heritage.org (дата звернення: 08.04.2023).

16. Рейтинг державних фондів. URL: https://www.fitchratings.com (дата звернення: 08.04.2023).

17. Статистика кількості та якості земельних ресурсів у світі. URL: http://www.iiasa.ac.at/ (дата звернення: 08.04.2023).

18. Статистика світових цін на газ, нафту. URL: https://www.eni.com/.

19. NGO: Energy Policy Group. Статистика з енергетичного ринку в СС, газринок. URL: http://energywatchgroup.org (дата звернення: 08.04.2023).

20. Тіньова економіка, офшори. URL: http://www.gfmtegrity.org/ (дата звернення: 08.04.2023).