Основи статистичного аналізу даних. Ряди даних

Інформатика

Рівень стандарту 10(11)

Урок 11

за підручником

Морзе Н.В. та ін.

2-ге видання, оновлене

Розділ 2
§ 10

Основи статистичного аналізу даних. Ряди даних

Пригадай

- основи роботи в табличному процесорі;
- використання майстра функцій табличного процесора.

Ти дізнаєшся

- що таке аналіз даних та які виділяють етапи аналізу даних;
- у чому суть статистичного підходу до опрацювання даних;
- як будують ряди даних;
- що є основними статистичними характеристиками вибірки;
- які функції можна використовувати для знаходження центральної тенденції в середовищі табличного процесора.

Розділ 2
§ 10

Що таке аналіз даних та які виділяють етапи аналізу даних?

Аналіз даних — розділ математики й інформатики, що займається розробкою методів опрацювання даних незалежно від їх природи.

Для аналізу даних потрібні знання предметної області та знання математики й статистики. Розуміння предметної області дає змогу визначити, які проблеми потребують першочергового вирішення. Знання математики й статистики дають змогу формалізувати рішення, перевести його в алгоритм та оцінити, яка ймовірність отримати результат; для цього використовують засоби комп'ютерної техніки.

Розділ 2
§ 10

Що таке аналіз даних та які виділяють етапи аналізу даних?

Розрізняють чотири етапи аналізу даних.

Розділ 2
§ 10

Що таке аналіз даних та які виділяють етапи аналізу даних?

Спочатку дані необхідно підготувати, тобто зібрати та відібрати ті, які потрібні для моделі опрацювання. Далі будується модель опрацювання й аналізуються її результати. Останній етап — це інтерпретація та презентація результатів. Тут потрібно продемонструвати питання, на яке шукали відповідь, які дані використовували та що отримали в результаті.

При збиранні даних використовують різні рівні їх виміру. Наприклад, розрізняють значення в певній точці (8; 3,6);.

можна розглядати різні інтервали	***та різні відношення***
([1,6] [0,15])	***(86 %, 14 %)***

Розділ 2
§ 10

У чому суть статистичного підходу до опрацювання даних?

Коли ми робимо виміри, то завжди існує ймовірність похибки. Багаторазове вимірювання та збереження при цьому відповідних результатів приводить до накопичення даних, які опрацьовують спеціальними методами, які вивчаються у статистиці. Такі дані називаються статистичними даними.

Статистичні дані — сукупність упорядкованих, класифікованих даних про деяке масове явище або процес.

Розділ 2
§ 10

У чому суть статистичного підходу до опрацювання даних?

Статистичні дані дають змогу не тільки охопити картину певного питання на даний час, а й планувати необхідні дії на майбутнє.

Так, статистичні дані про зайнятість населення дають можливість визначити:

яку кількість спеціалістів і якої кваліфікації слід готувати

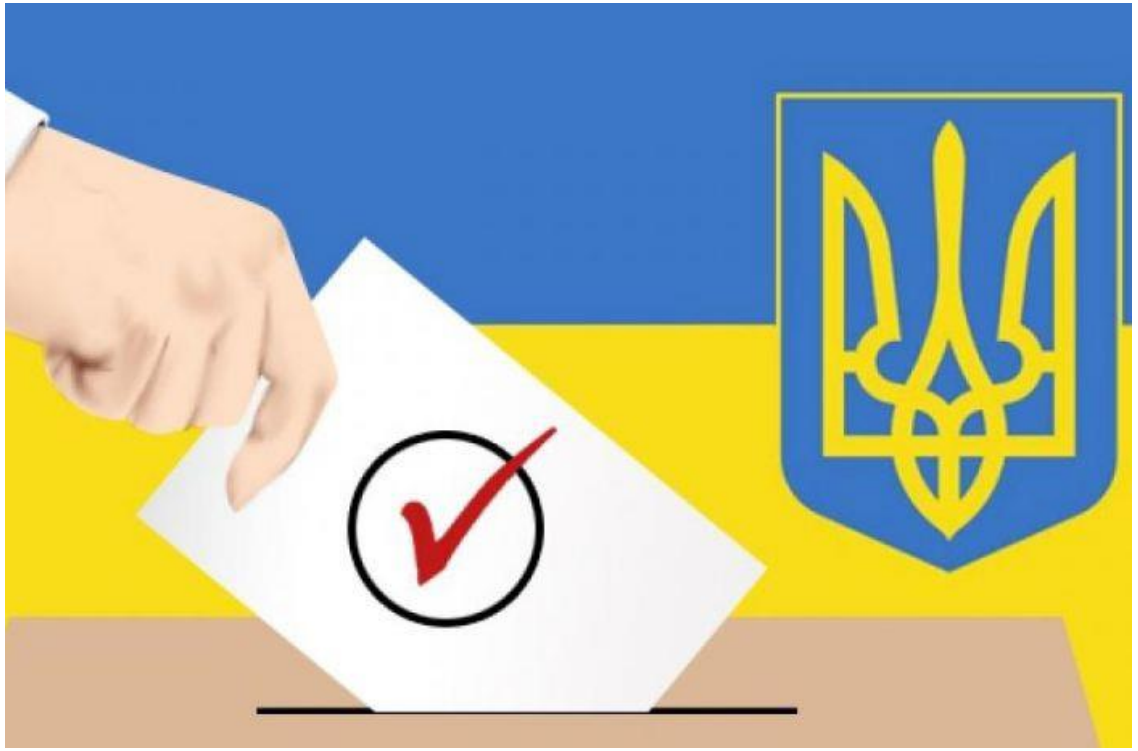
у якому регіоні варто споруджувати те чи інше підприємство.

Розділ 2
§ 10

У чому суть статистичного підходу до опрацювання даних?

Велику множину об'єктів, що є предметом статистичного дослідження, називають генеральною сукупністю.

Наприклад, якщо досліджуються передвиборчі вподобання, генеральною сукупністю може бути населення країни. Проте дослідник, як правило, не має змоги оперувати всією генеральною сукупністю.



Розділ 2
§ 10

У чому суть статистичного підходу до опрацювання даних?

Наприклад, опитати кожного громадянина країни нереально. Натомість досліджують вибірку — деяку множину об'єктів, вибраних з генеральної сукупності, і, проаналізувавши її, роблять висновки щодо властивостей генеральної сукупності загалом.

Так, дослідивши вподобання 10 000 виборців, можна зробити достатньо точні висновки щодо вподобань виборців усієї країни.

Розділ 2
§ 10

У чому суть статистичного підходу до опрацювання даних?

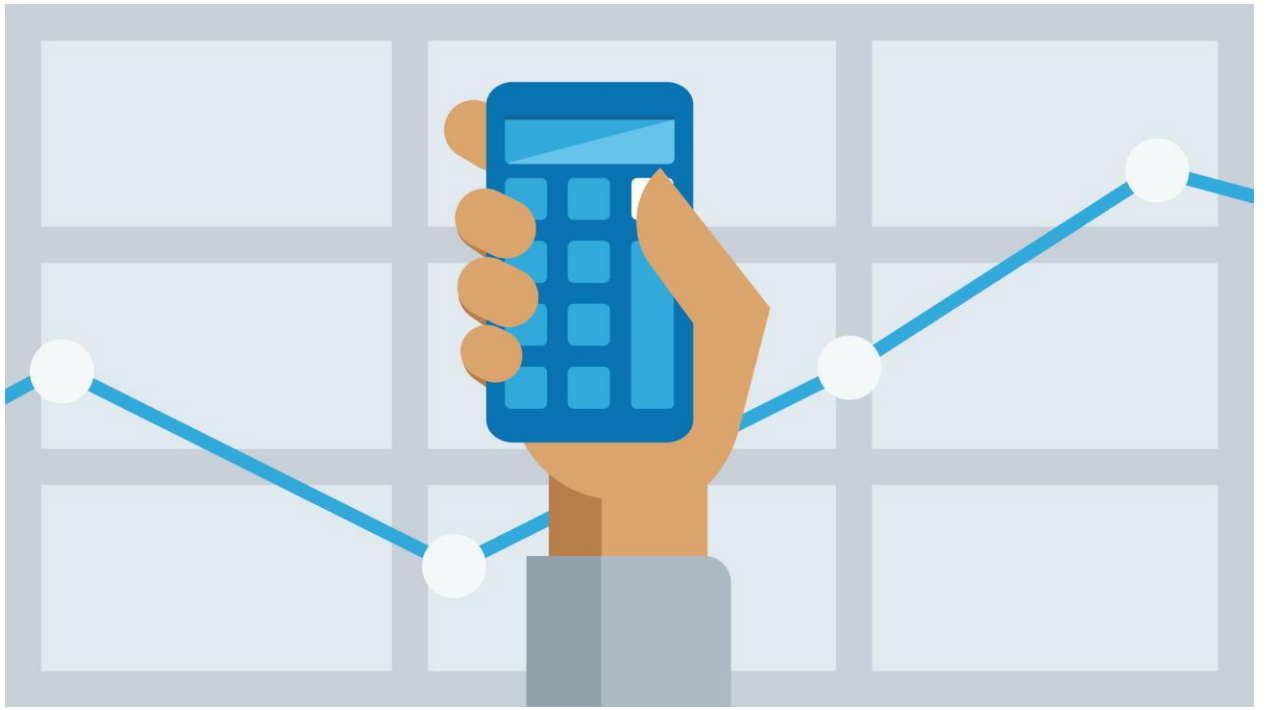
У заміні дослідження великої множини об'єктів дослідженням значно меншою її частиною та подальшому «поширенні» результатів дослідження на всю множину полягає сутність статистичного підходу до опрацювання даних.

Розділ 2
§ 10

Як будують ряди даних?

Маючи в розпорядженні дані статистичного спостереження, що характеризують те чи інше явище, перш за все необхідно їх упорядкувати, тобто надати характер системності.

Статистичний ряд розподілу — це впорядковані статистичні дані.



Розділ 2
§ 10

Як будують ряди даних?

Найпростішим видом статистичного ряду розподілу є ранжований ряд, тобто ряд чисел, що розташовані в порядку зростання чи спадання ознаки, яка змінюється.

Такий ряд не дає змоги судити про закономірності, закладені в розподілених даних:

біля якої величини групується більшість показників;	***які є відхилення від цієї величини;***	***яка загальна картина розподілу.***

Із цією метою дані групують, показуючи, як часто трапляються окремі спостереження в загальній їх кількості.

Розділ 2
§ 10

Як будують ряди даних?

Ряди розподілу одиниць сукупності за ознаками, що мають кількісний вираз, називаються варіаційними рядами. У таких рядах значення ознаки (варіанти) розташовані в порядку зростання або спадання.

Розділ 2
§ 10

Як будують ряди даних?

У варіаційному ряді розподілу розрізняють два елементи:

варіанта	***частота***
це окреме значення групувальної ознаки	***число, яке показує, скільки разів трапляється кожна варіанта***

Варіанта	Оцінка	1	2	3	4	5	6	7	8	9	10	11	12
Частота	Кількість учнів	0	0	0	1	1	2	1	2	4	3	2	1

Дискретний варіаційний ряд

Розділ 2
§ 10

Як будують ряди даних?

Таким чином, варіаційний ряд розподілу — це такий ряд, у якому варіанти розташовані в порядку зростання або спадання, вказані їх частоти або частки. Варіаційні ряди бувають:

дискретні	***інтервальні***
це такі ряди розподілу, в яких варіанта як величина кількісної ознаки може набувати тільки певного значення. Варіанти різняться між собою на одну чи кілька одиниць.	***такі ряди розподілу, в яких значення варіанти дано у вигляді інтервалів, тобто значення ознак можуть відрізнятися одне від одного на як завгодно малу величину. Можуть бути рівні й нерівні.***

Розділ 2
§ 10

Як будують ряди даних?

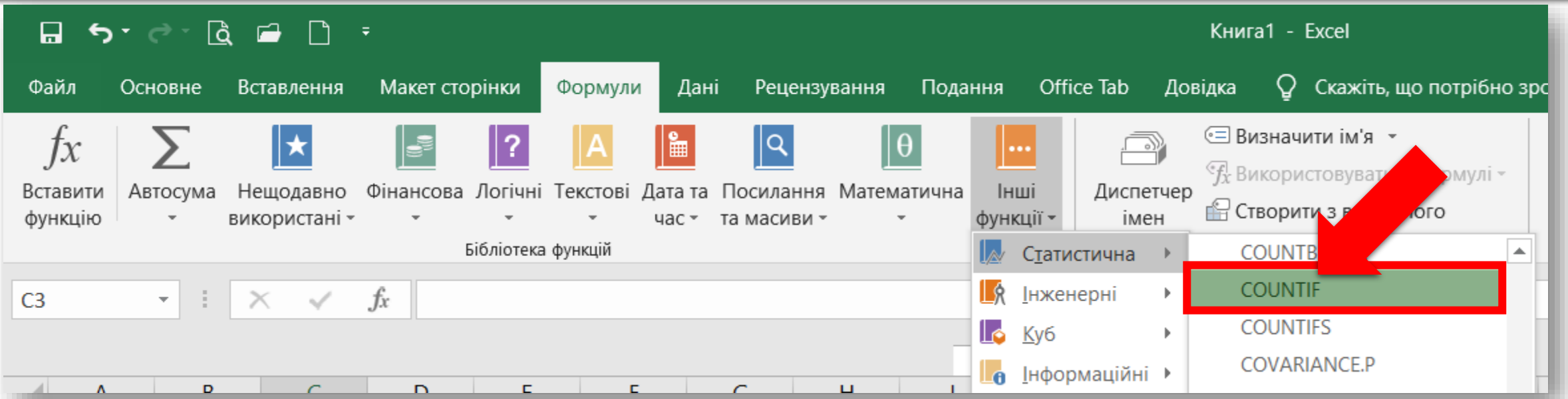
Наприклад, відомості про розподіл областей України із чисельністю населення на 1 грудня 2017 р., за даними ukrstat.org, можна подати інтервальним варіаційним рядом.

до 1 млн	від 1 млн до 1,5 млн	від 1,5 млн до 2 млн	від 2 млн до 2,5 млн	від 2,5 млн до 3 млн	понад 3 млн
2	13	3	2	3	2

Розділ 2
§ 10

Як будують ряди даних?

Для побудови дискретного ряду розподілу слід виписати всі можливі значення ознаки, а потім підрахувати, скільки разів кожне з них трапляється у вибірці — це будуть частоти. У Microsoft Excel для підрахунку частот можна застосувати функцію COUNTIF з категорії Статистичні.



Розділ 2
§ 10

Як будують ряди даних?

Для побудови за вибіркою x_1 , … , x_n ряду розподілу, що складається з m рівних інтервалів, необхідно виконати такі кроки:

1. Визначити найбільшу та найменшу варіанти — x_{min} та x_{max}

2. Визначити величину інтервалу

$$h = \frac{x_{max} - x_{min}}{m}$$

Розділ 2
§ 10

Як будують ряди даних?

Продовження...

3. Визначити межі інтервалів $[y_0, y_1], [y_1, y_2], \dots, [y_{m-1}, y_m]$, за формулами:

$$y_0 = x_{min};\ y_{i+1} = y + h, i = 0, \dots, m-1.$$

Тобто нижня межа першого інтервалу дорівнює найменшій варіанті, а кожна наступна межа більша за попередню на h.

Розділ 2
§ 10

Як будують ряди даних?

Продовження...

4. ***Підрахувати, скільки варіант потрапляє в кожний інтервал, — це й будуть частоти. У Microsoft Excel це можна зробити за допомогою функції FREQUENCY, яка має два аргументи:***

FREQUENCY(діапазон_ вибірки; діапазон_меж_ інтервалів)

діапазон, що містить вибірку

діапазон усіх меж інтервалів, за винятком y_0 та y_m (тобто всіх меж між інтервалами).

Розділ 2
§ 10

Як будують ряди даних?

Результатом функції буде набір частот, що відповідають кожному інтервалу. Ви вперше стикаєтеся з функцією, результатом якої є діапазон значень, а не окреме значення.

Її і вводити потрібно дещо інакше, ніж інші функції. А саме, слід виділити весь діапазон, де міститимуться результати, ввести формулу функції та натиснути клавіші:

Ctrl + Shift + Enter

Розділ 2
§ 10

Що є основними статистичними характеристиками вибірки?

Основними статистичними характеристиками вибірки є:

середнє	***мода***	***медіана***

Які ще називають мірами центральної тенденції. Вони показують загальні або типові характеристики розподілу даних за певною змінною.

Середнє, мода та медіана — це окремі значення, що представляють весь набір даних, типові для всіх значень у групі.

Розділ 2
§ 10

Що є основними статистичними характеристиками вибірки?

Розглянемо кожну з них.

Для обчислення середнього значення досить додати всі значення в розподілі й поділити на кількість спостережень.

Розділ 2
§ 10

Що є основними статистичними характеристиками вибірки?

Медіану *можна визначити як точку на ряді розподілу (впорядкований набір значень змінної для різних спостережень — наприклад, від найменшого до найбільшого значення) — до цієї точки розташована половина всіх значень, і після цієї точки — теж половина значень.*

Тобто медіана — це значення, що ділить упорядкований 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.

Розділ 2
§ 10

Що є основними статистичними характеристиками вибірки?

Мода *— це значення, яке найчастіше трапляється. Як правило, вона представляє найбільш типове значення. Наприклад, на сайті:*

radiotrek.rv.ua/news/yak-ukrayinci-nazivali-ditey-u-2022-roci-naypo-pulyarnishi-imena_301595.html

Дано перелік найпопулярніших імен дітей у II півріччі 2022 року. Ці імена є модою серед усіх імен. На моду ніколи не впливають екстремальні значення в розподілі, а впливають екстремальні частоти значень, наскільки часто те чи інше значення змінної трапляється в розподілі

Розділ 2
§ 10

Що є основними статистичними характеристиками вибірки?

Кожне з мір центральної тенденції має загальні правила для використання, переваги та обмеження.

Застосування	Переваги	Обмеження
Середнє арифметичне		
➢ «центр тяжіння» розподілу, і кожне значення робить внесок у визначення середнього значення, коли поширення значень є симетричними довкола центральної точки; ➢ коли потрібно знайти найстабільнішу міру центральної тенденції, використовують середнє арифметичне	➢ визначене однозначно, тому не виникає питань чи нерозуміння щодо його значення та суті; ➢ легко підрахувати та зрозуміти; ➢ ураховує всі значення розподілу	➢ на значення впливають екстремальні; наприклад, для знаходження середньої заробітної плати із сум 6000, 5000, 7000, 25 000 сума 25 000 суттєво збільшує середнє значення — 10 750, а при відкиданні найбільшого значення — середнім є сума 6000 грн; ➢ часом є таке значення, що відсутнє в розподілі або ж абсурдне. Наприклад, маємо 28, 31, та ЗО учнів у 5а, 56 та 5в класах якоїсь школи. Виходить, що середня кількість учнів у 5 класах школи — 29,(6). А так не буває

Розділ 2
§ 10

Що є основними статистичними характеристиками вибірки?

Продовження...

Застосування	*Переваги*	*Обмеження*

Медіана

➢ *для знаходження точної середньої точки, точку на «пів-дорозі» від найменшого значення до найбільшого;* ➢ *не враховує екстремальні значення;* ➢ *використовують, коли потрібно, щоб певні значення впливали на центральну тенденцію, але все, що про них відомо, — це те, що вони «нижчі» або «вищі» за медіану*	➢ *легко вирахувати та зрозуміти;* ➢ *для підрахунку не потрібні всі значення в розподілі;* ➢ *Екстремальні значення розподілу не впливають на результат*	➢ *значення не так вираховують, як знаходять (серед значень у розподілі);* ➢ *не враховує всі спостереження (значення для всіх спостережень);* ➢ *не можна робити алгебраїчні перетворення так само, як із середнім;* ➢ *потребує впорядкування значень або інтервалів у висхідному чи спадному порядку;* ➢ *часом може бути значення, відсутнє в самому розподілі*

Розділ 2
§ 10

Що є основними статистичними характеристиками вибірки?

Продовження...

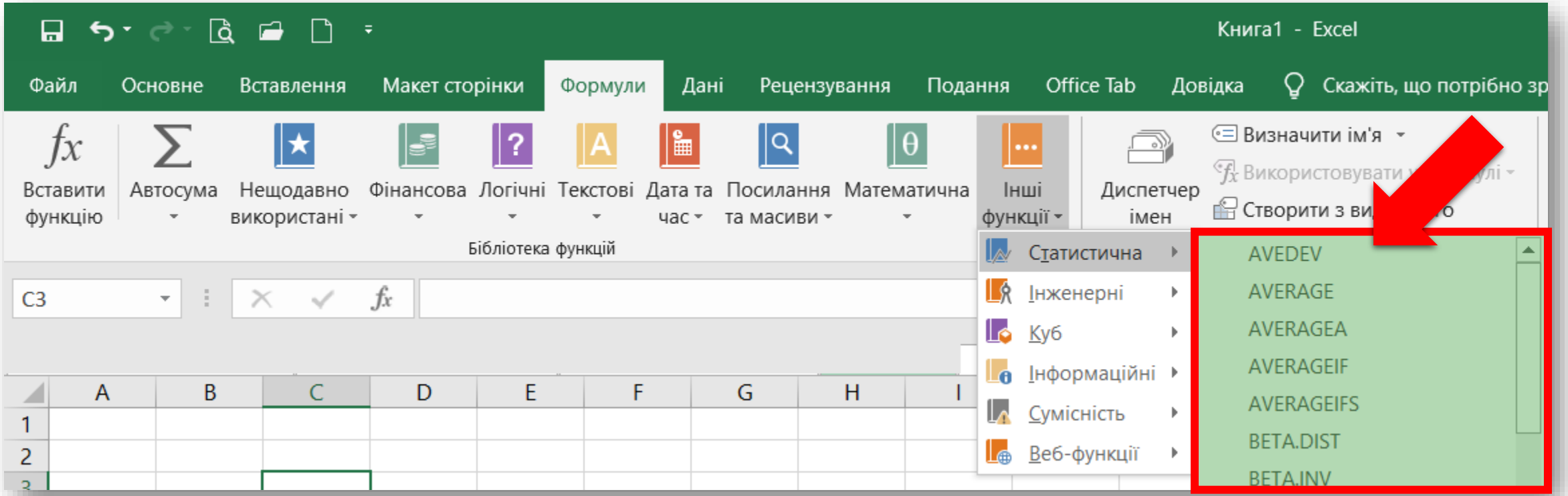
Застосування	*Переваги*	*Обмеження*
Мода		
➢ *коли потрібна швидка й приблизна міра центральної тенденції.* ➢ *для знаходження типового значення*	➢ *показує найбільш поширене значення в розподілі;* ➢ *можна аналізувати якісні дані;* ➢ *можна виявити, побудувавши графік розподілу чи стовпчасту діаграму*	➢ *не включає до визначення або розрахунку всі спостереження розподілу, а лише концентрацію частот;* ➢ *подальші алгебраїчні перетворення неможливі — на відміну від середнього*

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

Розділ 2
§ 10

Які функції можна використовувати для знаходження центральної тенденції

У табличному процесорі центральна тенденція представлена функціями з категорії Статистичні, та майже всі вони призначені для обчислення узагальнювальних статистичних характеристик вибірки.



Розділ 2
§ 10

Які функції можна використовувати для знаходження центральної тенденції

Статистичні функції, що розглядаються нижче (крім функції RANK), можуть мати декілька аргументів, які мають бути:

числами	***масивами***	***посиланнями на діапазони клітинок, що містять числа***

Якщо до діапазону-аргументу функції входять клітинки, які містять текст чи логічні значення або є порожніми, то вони ігноруються; але клітинки, що містять нульові значення, враховуються. Коли потрібно обчислити певну статистичну характеристику вибірки, діапазон, що містить елементи вибірки, слід зробити аргументом функції.

Розділ 2
§ 10

Які функції можна використовувати для знаходження центральної тенденції

Ознайомимось із кількома найважливішими функціями категорії Статистичні:

AVERAGE	***обчислює середнє значення***
MAX	***обчислює максимальне значення***
MIN	***обчислює мінімальне значення***
MEDIAN	***повертає медіану***
MODE	***повертає моду***

Розділ 2
§ 10

Які функції можна використовувати для знаходження центральної тенденції

RANK — *повертає ранг числа у списку чисел, тобто його номер у впорядкованій послідовності чисел із вказаного діапазону*

RANK (*число*; *посилання*; *порядок*)

- *число, для якого визначається ранг*
- *масив або посилання на список чисел*
- *аргумент, який визначає спосіб упорядкування*

Якщо цей аргумент відсутній або дорівнює нулю, то найбільше число має ранг 1. Якщо цей аргумент дорівнює будь-якому ненульовому числу, то ранг 1 має найменше число.

Розділ 2
§ 10

Які функції можна використовувати для знаходження центральної тенденції

Примітка. ***Функція RANK призначає повторюваним числам однаковий ранг.***

Проте наявність повторюваних чисел впливає на ранги наступних чисел. Наприклад, якщо у списку цілих чисел, відсортованих за зростанням, двічі трапляється число 10 з рангом 5, число 11 матиме ранг 7 і жодне із чисел не матиме рангу 6.

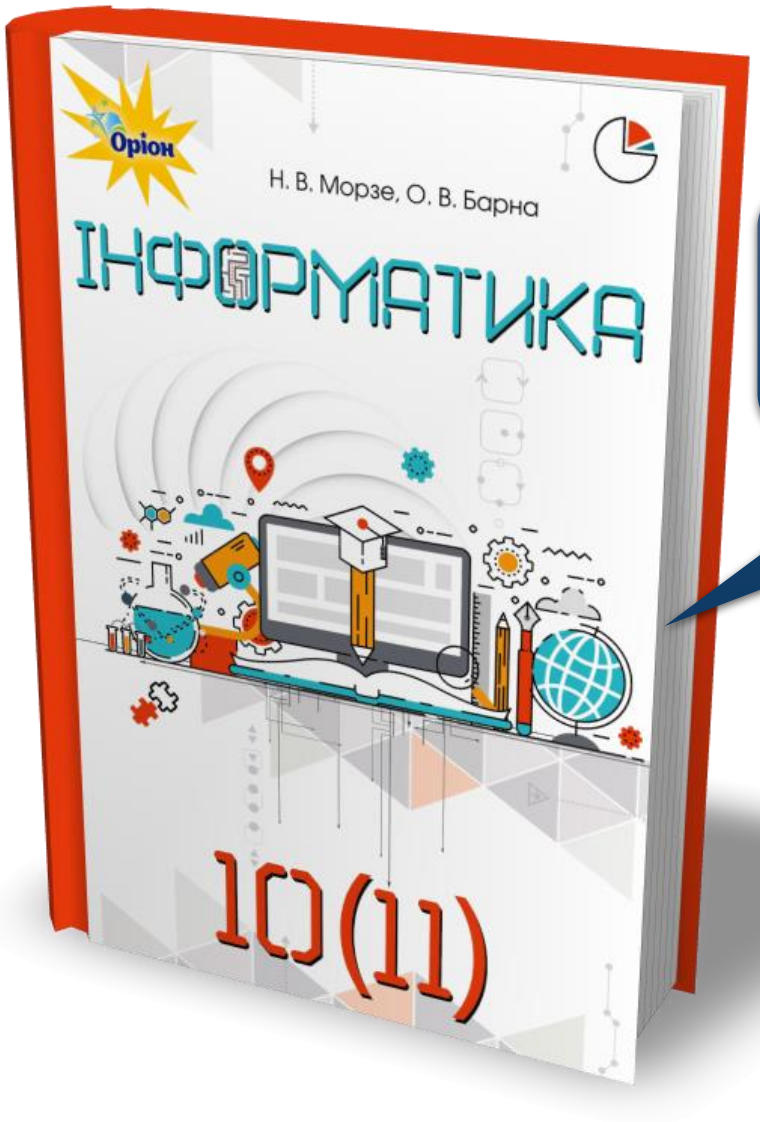
Розділ 2
§ 10

Дайте відповіді на запитання

1. *Що таке аналіз даних та які виділяють етапи аналізу даних?*
2. *Як статистика допомагає аналізу даних?*
3. *У чому суть статистичного підходу до опрацювання даних?*
4. *Як будують ряди даних?*
5. *Що таке вибірка?*
6. *Що є основними статистичними характеристиками вибірки?*

Розділ 2
§ 10

Домашнє завдання



Проаналізувати
§ 10, с. 91-100

Розділ 2
§ 10

Працюємо за комп’ютером

Сторінка 92-100

Інформатика

Рівень стандарту 10(11)

Урок 11

за підручником

Морзе Н.В. та ін.

Дякую за увагу!

2-ге видання, оновлене