Основи статистичного аналізу даних. Ряди даних. Обчислення основних статистичних характеристик вибірки

Основи статистичного аналізу даних. Ряди даних

Статистика - (лат status — стан) наука, що вивчає методи отримання, опрацювання й аналізу даних, які характеризують масові явища.
.

Етапи статистичних досліджень

- Статистичні спостереження.
- Первинне узагальнення і групування статистичних даних.
- Оцінка об'єкта аналізу.
- Комп'ютерний аналіз первинних і узагальнених розширених статистичних даних.
- Комп'ютерне прогнозування за обраними найбільш важливими напрямками.
- Узагальнений аналіз отриманих результатів та перевірка їх на достовірність за статистичними критеріями.

Ряди даних

Для аналізу створюють певну вибірку об’єктів дослідження, тобто з усієї множини об’єктів дослідження відбирають певну кількість і на ній проводять дослідження. **Вибірка** (або вибіркова сукупність) — це множина об’єктів, за допомогою певної процедури вибраних із генеральної сукупності для участі в дослідженні. **Варіанта** - значення величини у вибірці.

Чим більше така вибірка, тим точніше буде проведено аналіз і зроблено відповідні висновки. Тобто вибірка повинна бути масовою.

Вибірка даних має бути репрезентативною (франц representatif — показовий, характерний, типовий). Дані, отримані з дослідженої вибірки, найчастіше заносять у таблицю. Така форма подання даних з вибірки зручна для їх аналізу та прогнозів. Дані з кожного рядка і стовпця такої таблиці утворюють **ряди даних**.

Наведемо кілька прикладів вибірок і рядів даних. У таблиці подано результати виступів команди учнівства України на міжнародних олімпіадах з інформатики з 2005 по 2017 рік. Тут вибіркою є вказані в таблиці роки, а рядами даних — загальна кількість медалей у ці роки, а також кількість золотих, срібних і бронзових медалей у вказані роки.

Рік	*Кількість медалей*	*Золоті*	*Срібні*	*Бронзові*
2005	4	2	1	1
2006	4	1	2	1
2007	4	1	2	1
2008	3	0	1	2
2009	4	1	1	2
2010	3	0	1	2
2011	3	0	1	2
2012	4	1	1	2
2013	4	0	1	3
2014	3	0	1	2
2015	4	0	3	1
2017	3	1	2	0

Результати виступу команди учнівства України на міжнародних олімпіадах з інформатики

Упорядкований розподіл одиниць сукупності на групи за кількісною ознакою називають варіаційним рядом. Побудувати варіаційний ряд - означає упорядкувати кількісний розподіл одиниць сукупності за значеннями ознаки, а потім підрахувати число одиниць сукупності з цими значеннями (побудувати групову таблицю). **Варіаційні ряди бувають**: дискретними та інтервальними.

Дискретний варіаційний ряд

Дискретний ряд - це такий варіаційний ряд, в основу побудови якого покладено ознаки з переривчастою зміною (дискретні ознаки). До останніх можна віднести тарифний розряд, кількість дітей в сім'ї, число працівників на підприємстві і т.д. Ці ознаки можуть приймати лише кінцеве число певних значень.

Дискретний варіаційний ряд являє таблицю, яка складається з двох граф. У першій графі вказується конкретне значення ознаки, а в другій - число одиниць сукупності з певним значенням ознаки.

Приклад:

Є список оцінок, отриманих студентами на іспитах: 3; 4; 3; 5; 4; 2; 2; 4; 4; 3; 5; 2; 4; 5; 4; 3; 4; 3; 3; 4; 4; 2; 2; 5; 5; 4; 5; 2; 3; 4; 4; 3; 4; 5; 2; 5; 5; 4; 3; 3; 4; 2; 4; 4; 5; 4; 3; 5; 3; 5; 4; 4; 5; 4; 4; 5; 4; 5; 5; 5.

Побудувати дискретний варіаційний ряд.
Розв'язання:
1) Впорядкуємо одиниці спостереження за зростанням досліджуваного значення ознаки: 2; 2; 2; 2; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 4; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5; 5.

2) Визначимо всі можливі значення ознаки xi, впорядкуємо їх за зростанням: У даному прикладі всі оцінки можна розділити на чотири групи з наступними значеннями: 2; 3; 4; 5.

3) Підрахуємо скільки разів зустрічається кожне значення ознаки у досліджуваній сукупності, тобто визначимо частоту кожного значення ознаки fi. Частота - число, що показує, скільки разів зустрічається кожна варіанта.
Відносна частота - відношення частоти випадків даного значення до загальної суми частот.
Сума всіх частот ряду дорівнює кількості елементів у досліджуваній сукупності.

Для нашого прикладу:

- оцінка 2 зустрічається - 8 разів,
- оцінка 3 зустрічається - 12 разів,
- оцінка 4 зустрічається - 23 рази,
- оцінка 5 зустрічається - 17 разів.

Всього 60 оцінок.
4) Запишемо отримані дані в таблицю з двох рядків (стовпців) - xi і fi. На підставі цих даних побудуємо дискретний варіаційний ряд:

x_i (оцінка)	f_i (к-ть студентів з такою оцінкою)
2	8
3	12
4	23
5	17
Разом	60

З метою створення візуального відображення статистичної інформації користуються різними графіками. Найпоширеніші види графічного відображення статистичної інформації — полігони частот. Графічне зображення варіаційних рядів за допомогою полігона допомагає отримати наочне уявлення про закономірності про можливі зміни спостережуваних значень.

Полігон, як правило, використовують для відображення дискретного варіаційного ряду.

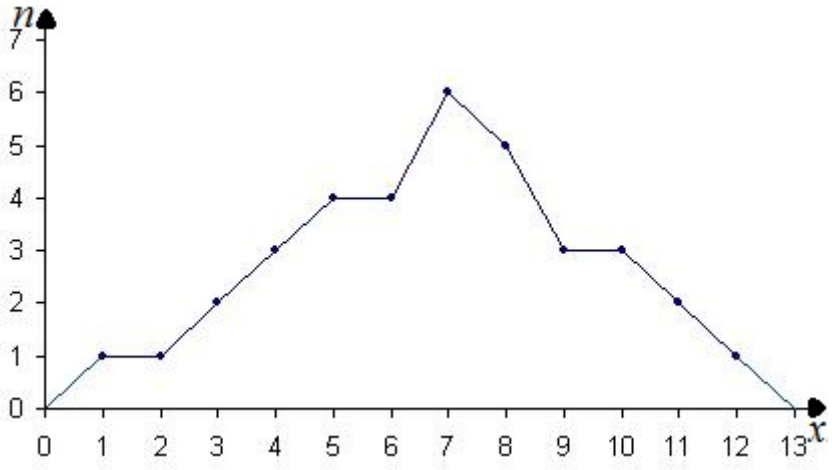
Приклад:

Навчальні досягнення учнів деякого класу з інформатики характеризуються даними, наведеними у таблиці:

Кількість балів *x*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Число учнів *n*	1	1	2	3	4	4	6	5	3	3	2	1

Побудувати полігон частот.

Розв'язання: Будуємо точки, спираючись на дані з таблиці. Отримані точки з'єднуємо відрізками. Зверніть увагу на точки (0; 0) та (13; 0), що розташовані на осі абсцис і мають своїми абсциссами числа, на одиницю менше та більше, ніж відповідно абсциси найлівішої та найправішої точок. Полігон частот виглядатиме наступним чином:



Інтервальний варіаційний ряд

Якщо ознака має безперервну зміну (розмір доходу, стаж роботи, вартість основних фондів підприємства і т.д., які в певних межах можуть приймати будь-які значення), то для цієї ознаки потрібно будувати інтервальний варіаційний ряд.

Групова таблиця тут також має дві графи. У першій вказується значення ознаки в інтервалі «від - до» (варіанти), у другій - число одиниць, що входять до інтервалу (частота).

Частота - число повторень окремого випадку значень ознаки, позначається fi, а сума частот, що дорівнює обсягу досліджуваної сукупності, позначається $\sum_{i=1}^{k} f_i$, де k - число варіантів значень вибірки.

Алгоритм побудови інтервального варіаційного ряду:

- визначити кількість інтервалів для побудови інтервального варіаційного ряду;
- знайти ширину інтервалу (крок) h = (xmax - xmin) / k, де k - число варіантів значень вибірки;
- визначити межі кожного інтервалу в порядку зростання;
- підрахувати число даних, що потрапили в кожний з інтервалів.

Приклад:

За результатами аналізу вугілля, відібраного безповторним відбором, отримані наступні

результати щодо його зольності : 18, 16, 18, 21, 19, 17, 18, 21, 14, 18, 16, 12, 19, 23, 17, 18, 15, 20, 19, 17, 21, 16, 20, 13, 19, 14, 20, 15,18, 20, 17, 19, 16, 18, 13, 15, 17, 24, 16, 14.

Необхідно побудувати інтервальний варіаційний ряд, поділивши дані на шість рівних інтервалів.

Розв'язання:

- k = 6;
- h = (24 - 12) / 6 = 2;
- 12 - 14
    1. 14 - 16
    2. 16 - 18
    3. 18 - 20
    4. 20 - 22
    5. 22 - 24
- 6; 11; 17; 16; 7; 2.

x_i	f_i
12 - 14	6
14 - 16	11
16 - 18	17
18 - 20	16
20 - 22	7
22 -24	2

Розглянемо деякі статистичні характеристики вибірки: **середнє арифметичне, стандартне відхилення, мода і медіана**.

Середнє арифметичне

Ви знаєте, що середнім арифметичним n чисел називається сума цих чисел, поділена на число **n**.

Так, можна знайти середнє арифметичне врожайності соняшнику в Україні за 2006–2015 роки, використовуючи, наприклад, табличний процесор.

Для обчислення середнього арифметичного в табличному процесорі можна використати відому вам функцію AVERAGE(СРЗНАЧ) (англ. average— середній). Нагадаємо, що аргумен тами цієї функції може бути діапазон клі тинок, список клітинок, а також їх комбінації, наприкладAVERAGE (B2:D5; F4; E7). На малюнку було наведено приклад обчислення середньої врожайності соняшнику за 2006–2015 роки і формулу для її обчислення =AVERAGE (C3:C12).

Обчислене в наведеному прикладі середнє арифметичне визначає, яка б була врожайність кожного року (1,67 т/га), якщо вона щороку була б однаковою. Аналогічно середнє арифметичне будь-якого ряду даних визначає, які б були значення у цьому ряді, якщо б вони всі були однакові.

Зазначимо, що не для всіх рядів даних середнє арифметичне є показовою характеристикою самого цього ряду. Наприклад, для ряду даних 2,5; 2,8; 2,3; 2,55; 2,47, у якому дані незначно відрізняються одне від одного, середнє арифметичне дорівнює 2,524, що незначно відрізняється від усіх членів цього ряду, а значить, достатньо показово характеризує весь цей ряд даних. А для ряду 4,7; 6,2; 5,1; 12,4; 14,1, у якому дані значно відрізняються одне від одного, середнє арифметичне дорівнює 8,5, що значно відрізняється від усіх членів цього ряду, а значить, недостатньо показово характеризує весь цей ряд даних.

Стандартне відхилення

Для визначення, наскільки показово середнє арифметичне ряду даних характеризує весь ряд даних, можна використати таку характеристику ряду даних, як **стандартне відхилення**. Стандартне відхилення характеризує, наскільки широко розташовані значення ряду даних відносно їх середнього арифметичного.

Автоматизувати обчислення стандартного відхилення в табличному процесорі можна, використавши функцію STDEV.P (англ. standard deviation — стандартне відхилення) (для версії нижче 2010 — STDEVP).

Мода

Ще однією характеристикою ряду даних є мода. **Мода** — це значення в ряді даних, яке повторюється найчастіше. Таке значення є показовим, наприклад, під час дослідження цін на ринку (ціна, яка трапляється найчастіше), під час дослідження попиту взуття, одягу (розміри, які купують найбільше) та ін

У розглянутому вище прикладі мода кількостей медалей, які вибороло учнівство України на міжнародних олімпіадах з інформатики за 2005–2017 роки, дорівнює 4 (тому що найчастіше в ці роки команда нашої країни завойовувала 4 медалі), мода кількостей золотих медалей — 0, мода кількостей срібних медалей — 1, мода кількостей бронзових медалей — 2.

Якщо в ряді даних два або більше значень повторюються найбільшу кількість разів, то кожне з них вважається модою ряда даних Так, наприклад, у ряді даних 2, 3, 3, 2, 1 модою є і число 2, і число 3.

У табличному процесорі є спеціальна функція для обчислення моди ряду даних, якщо вона одна — MODE.SNGL (англ. mode single — мода одинарна) (для версії Excel нижче 2010 і для LibreOffice Calc — MODE) Аргументами цієї функції може бути діапазон клітинок, список клітинок, а також їх комбінації, наприклад MODE.SNGL (B2:D5; F4; E7).

На малюнку наведено приклад обчислення моди для кількостей завойованих медалей і формула для її обчислення: =MODE.SNGL (E6:E17).

Медіана

Розглянемо ще одну характеристику ряду даних — медіану.

Медіаною впорядкованого ряду даних називається значення, яке поділяє ряд даних на дві рівні частини, тобто зліва і справа від цього значення знаходиться однакова кількість членів упорядкованого ряду даних.

Якщо у впорядкованому ряді даних непарна кількість членів, то медіана такого ряду даних дорівнює значенню його середнього члена, а якщо в такому ряді даних парна кількість членів, то його медіана обчислюється як середнє арифметичне значень двох середніх членів.

Наприклад, для ряду даних 2; 3; 5; 6; 7 медіана дорівнює 5, для ряду даних 2; 3; 5; 6: 7; 9 медіана дорівнює (5 + 6) : 2 = 5,5, а для ряду даних 2; 2; 4; 4; 4; 5; 6 медіана дорівнює 4.

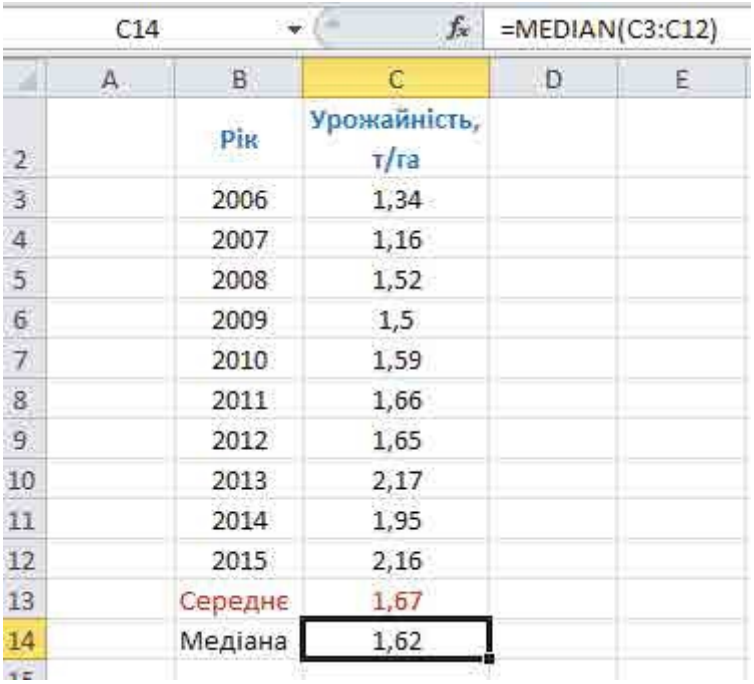
Медіана використовується, наприклад, для визначення місця побудови шкіл, дитячих садочків, магазинів, підприємств побуту тощо Потрібно визначити ряд відстаней, які слід подолати мешканцям певної місцевості до цього закладу, і побудувати його в точці, яка визначається медіаною цього ряду.

У табличному процесорі є спеціальна функція для обчислення медіани ряду даних — MEDIAN (англ. median — середній). Аргументами цієї функції може бути діапазон клітинок, список клітинок, а також їх комбінації, наприклад MEDIAN(B2:D5; F4; E7).

На малюнку наведено приклад обчислення медіани ряду даних урожайності соняшнику з використанням табличного процесора за формулою =MEDIAN(C3:C12).

X1	2,5	4,7
X2	2,8	6,2
X3	2,3	5,1
X4	2,55	12,4
X5	2,47	14,1
Середнє	2,524	8,5
Стандартне відхилення	0,1615673	3,946137
Медіана	2,5	6,2

Звертаємо вашу увагу, що в електронній таблиці для знаходження медіани ряд даних не обов’язково має бути впорядкований. Табличний процесор спочатку впорядковує ряд даних, а потім визначає його медіану.



C14 =MEDIAN(C3:C12)

	A	B	C	D	E
2		Рік	Урожайність, т/га		
3		2006	1,34		
4		2007	1,16		
5		2008	1,52		
6		2009	1,5		
7		2010	1,59		
8		2011	1,66		
9		2012	1,65		
10		2013	2,17		
11		2014	1,95		
12		2015	2,16		
13		Середнє	1,67		
14		Медіана	1,62		

Зазначимо, що коли члени ряду даних незначно відрізняються одне від одного, то і середнє арифметичне, і медіана більш показово характеризують весь цей ряд. А якщо члени ряду даних значно відрізняються одне від одного, то медіана більш показово характеризує весь цей ряд даних, ніж середнє арифметичне.

Джерела

ivanytskyi.blogspot.com