**Заняття №3**

**Методи статистичного опису результатів спостережень**

*Означення 1*. Сукупність спостерігаємих випадкових величин  будемо називати **вибіркою**. Реалізацію вибірки  позначимо через . Якщо всі елементи вибірки  є незалежними і однаково розподілені, як деяка випадкова величина , то таку вибірку називають вибіркою з **генеральної сукупності** з розподілом .

*Означення 2*. Кількість спостережень над випадковою величиною  називається **об’ємом вибірки**.

*Означення 3*. Нехай  - вибірка з генеральної сукупності з розподілом  і  - значення , що спостерігались. Кожній реалізації  вибірки  можна поставити у відповідність упорядковану послідовність

,

де,  - друге за величиною значення з  і т.д.; . Позначимо через  випадкову величину, яка для кожної реалізації вибірки  набуває значення 

За вибіркою  визначимо нову послідовність випадкових величин , які називаються **порядковими статистиками** вибірки. При цьому  - -та порядкова статистика, а  - мінімальна та максимальне значення вибірки відповідно.

З визначення порядкових статистик випливає, що вони задовольняють нерівності

. (1)

Послідовність (1) називається **варіаційним рядом** вибірки.

Зазначимо, що різниця між найбільшим та найменшим елементами реалізації вибірки називається **розмахом вибірки**

.

*Означення 4*. Нехай у вибірці об’єму  елемент  зустрічається  разів, тоді число  називається **частотою** елемента . Сукупність пар  називається **статистичним рядом** і записується у вигляді таблиці

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | .... |  |
|  |  | .... |  |

Зазначимо, що  - це кількість різних (без повторень) елементів вибірки. При цьому очевидними є такі співвідношення

1) ,

2) .

**Приклад 1.**

Записати у вигляді варіаційного і статистичного рядів наступну вибірку:

5, 3, 7, 10, 5, 5, 2, 10, 7, 2, 7, 7, 4, 2, 4. Вказати значення для  та .

1. Варіаційний ряд вибірки

.

1. Статистичний ряд

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2 | 3 | 4 | 5 | 7 | 10 |
|  | 3 | 1 | 2 | 3 | 4 | 2 |

3) 

**Групування даних**

Якщо дані, які ми спостерігаємо відповідають випадковій величині абсолютно неперервного типу, то для подальшої їх обробки може знадобитися групована вибірка. Запишемо у вигляді алгоритму побудову групованої вибірки.

**1. Перший крок**.

Визначимо кількість інтервалів групування. Для цього застосуємо формулу Стерджесса:

.

На практиці цю формулу зручно використовувати у такому вигляді

.

**2. Другий крок**.

Визначення довжини інтервалу групування:

.

**3. Третій крок.**

За початок першого інтервалу доцільно взяти таке число

,

Кінець першого інтервалу та початок другого інтервалу визначається за такою формулою

.

Продовжуючи далі, будемо мати:



У результаті роботи алгоритму, одержимо таблицю

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Інтервал |  |  | ..... |  |
| Частота |  |  | ..... |  |

Частота  визначається, як кількість елементів початкової вибірки, що потрапляють в -тий інтервал.

Зазначимо, що даний алгоритм носить рекомендаційний характер і відповідно існують інші підходи до групування даних.

Важливою є задача переходу від групованої вибірки до групованого статистичного ряду. При цьому під групованим статистичним рядом будемо розуміти сукупність пар  де ,  - відповідно середина та частота -го інтервалу. В результаті одержимо таблицю

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | .... |  |
|  |  | .... |  |

З цією таблицею можна працювати як зі звичайним статистичним рядом.

**Приклад 2.**

У таблиці задано вибірку, яка відповідає доходам мешканців містечка. Потрібно за згаданим вище алгоритмом згрупувати ці дані.

|  |  |  |  |
| --- | --- | --- | --- |
| № | Дохід | № | Дохід |
| 1 | 3820 | 13 | 6660 |
| 2 | 9470 | 14 | 5490 |
| 3 | 3490 | 15 | 5980 |
| 4 | 7790 | 16 | 6250 |
| 5 | 4210 | 17 | 8390 |
| 6 | 3870 | 18 | 3630 |
| 7 | 4490 | 19 | 6090 |
| 8 | 9620 | 20 | 10450 |
| 9 | 6200 | 21 | 6800 |
| 10 | 6350 | 22 | 6470 |
| 11 | 7430 | 23 | 9160 |
| 12 | 7670 | 24 | 5110 |

*Розв’язання*.

Визначаємо кількість інтервалів групування

.

Визначаємо довжину інтервалу групування



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Інтервал | [2620, 4360) | [4360, 6100) | [6100, 7840) | [7840, 9580) | [9580,11320) |
|  | 5 | 5 | 9 | 3 | 2 |

**Емпірична функція розподілу**

Визначимо для кожного дійсного  випадкову величину , яка дорівнює кількості елементів вибірки , значення яких не перевищують :

,

де  - кількість елементів скінченної множини. Функція, яка задається рівністю , називається **емпіричною функцією розподілу**. Функцію розподілу  випадкової величини , що спостерігається, називається **теоретичною функцією розподілу**.

**Теорема (В.І. Глівенко, 1933)** Нехай  - емпірична функція розподілу, яка побудована за вибіркою ,  - відповідна теоретична функція розподілу. Тоді

.

Емпірична функція розподілу будується наступним чином:

1) Якщо всі елементи вибірки є різними, тобто , тоді

 .

2) Якщо дані представлені у вигляді статистичного ряду

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | .... |  |
|  |  | .... |  |



**Приклад 3.**

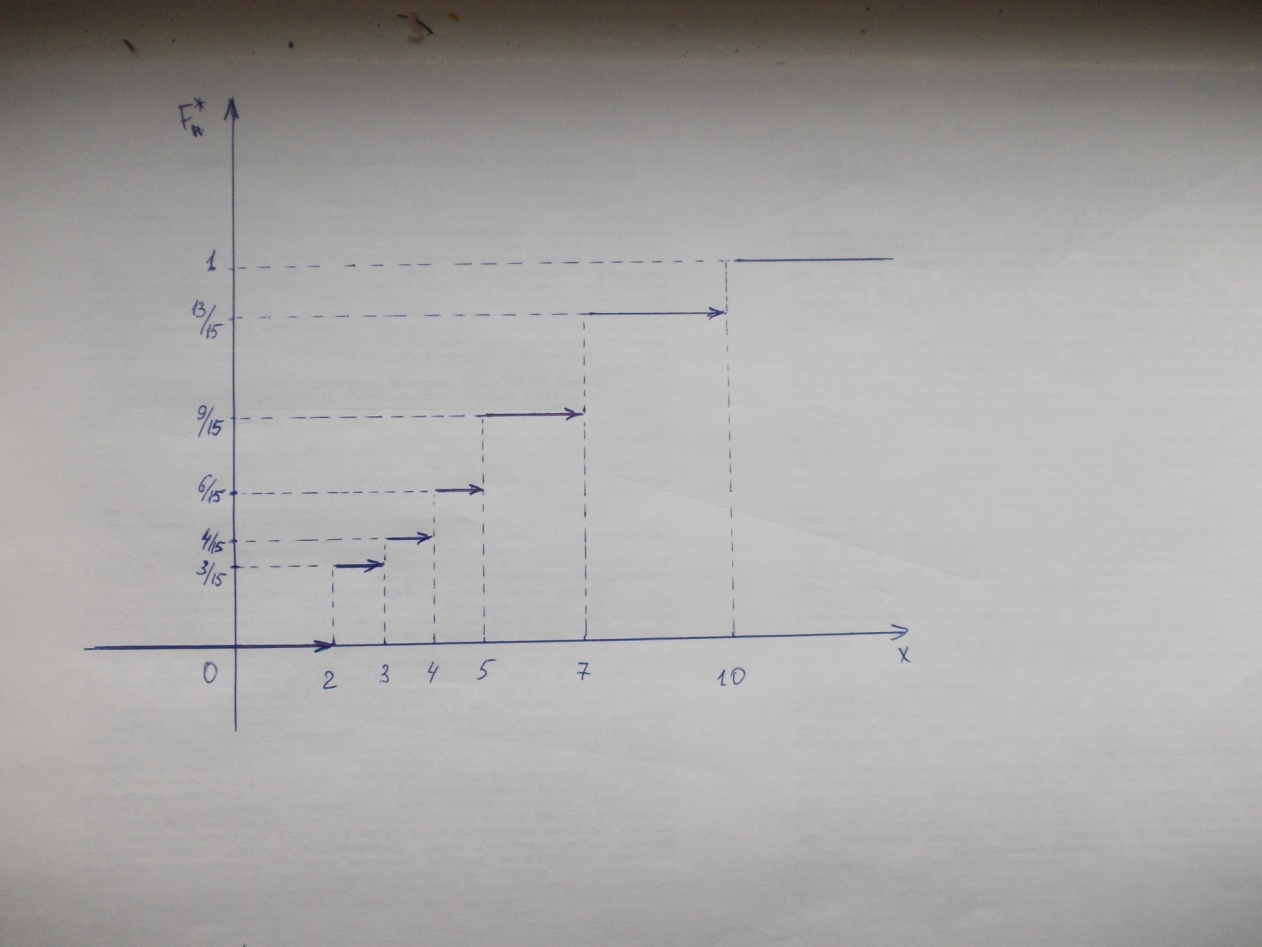
За вибіркою представленою у вигляді статистичного ряду

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2 | 3 | 4 | 5 | 7 | 10 |
|  | 3 | 1 | 2 | 3 | 4 | 2 |

побудувати емпіричну функцію розподілу і оцінити ймовірність того, що наступне спостереження потрапить до інтервалу .

*Розв’язання*.

Розв’язок представимо у вигляді графіка, враховуючи, що об’єм вибірки 

**

Щоб оцінити за допомогою емпіричної функції розподілу ймовірність того, що наступне спостереження потрапить до інтервалу , скористаємось властивістю функції розподілу

.

**Зауваження.** Як видно з прикладу, емпірична функція розподілу має вигляд східчастої лінії і у точках, де відбуваються стрибки функції є невизначеною похідна, а отже ми не зможемо оцінити щільність функції розподілу.

**Гістограма та полігон частот**

Гістограма та полігон частот - це статистичні аналоги щільності, які будуються тільки для групованих даних, що відповідають розподілам абсолютно неперервного типу.

1) Гістограма: на кожному інтервалі групування будується прямокутник з висотою

,

де  - об’єм вибірки,  - частота -го інтервалу,  - довжина -го інтервалу.

**Твердження.** Якщо об’єм вибірки  є великим, тоді гістограма  прямує до щільності теоретичної функції розподілу.

2) Полігон частот – це ламана лінія, яка з’єднує відрізками прямих наступні точки  , де  - середина -го інтервалу.

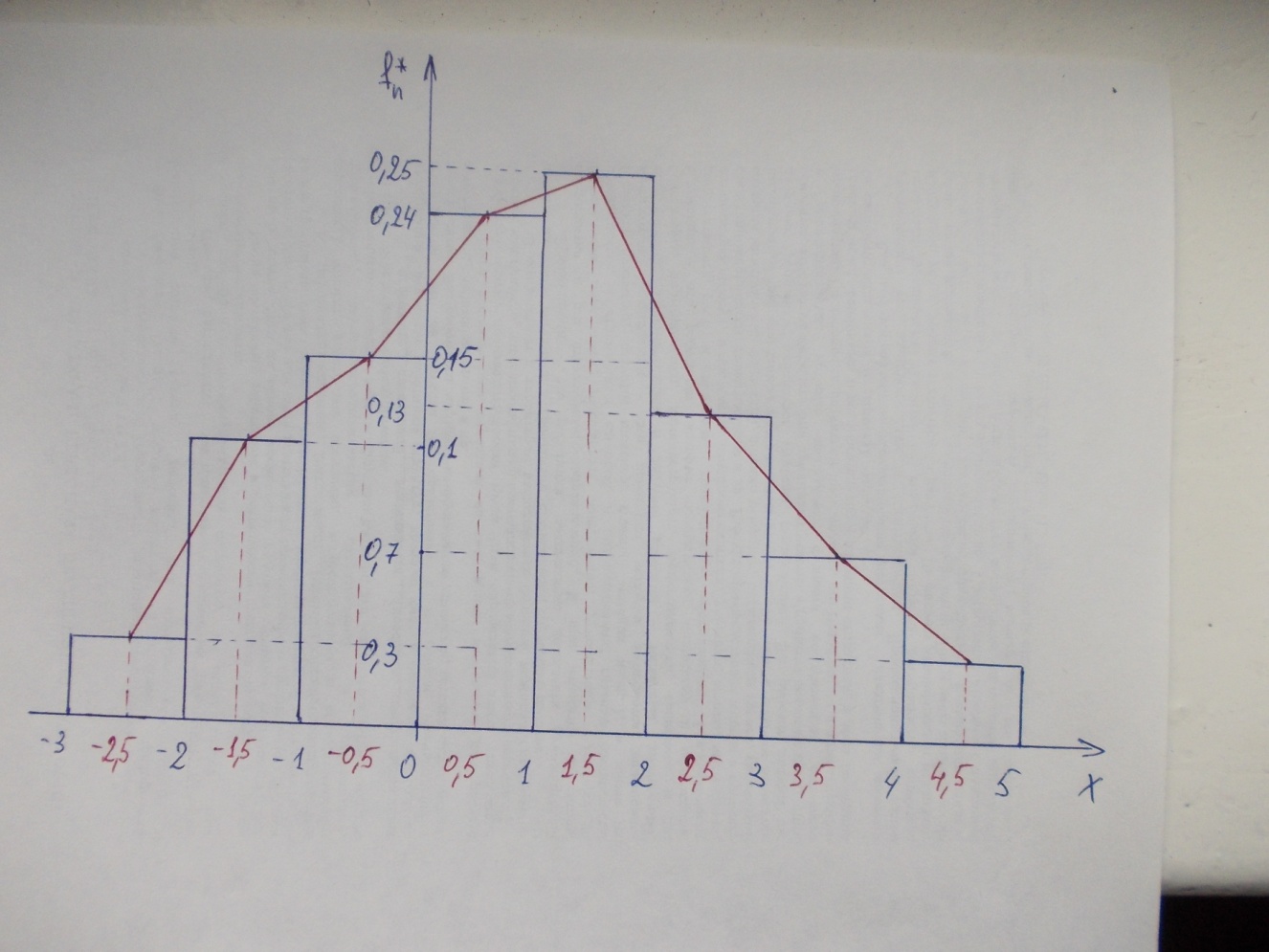
**Приклад 4.**

За наступною вибіркою побудувати гістограму та полігон частот

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Інтервал | [-3, -2) | [-2, -1) | [-1, 0) | [0, 1) | [1, 2) | [2, 3) | [3, 4) | [4, 5) |
|  | 3 | 10 | 15 | 24 | 25 | 13 | 7 | 3 |

*Розв’язання*.

Як і в попередньому прикладі, розв’язок представимо у графічному вигляді



На цьому малюнку синім кольором представлена гістограма, а червоним – полігон частот.

**Вибіркова медіана та вибіркові квантилі**

Розглянемо варіаційний ряд , сформований з порядкових статистик.

Нехай . Для неперервної функції розподілу  **теоретичним -квантилем**  називається мінімальний розв’язок рівняння



Якщо функція  строго монотонна, то це рівняння має єдиний корінь.

**Вибірковим -квантилем ** будемо називати порядкову статистику

****

Тут  - елемент вибірки, лівіше якого знаходиться частка  спостережень, при цьому  - порядкова статистика з максимальним номером, що задовольняє зазначену нерівність. Отже  можна розглядати як статистичний аналог .

**Вибіркова медіана** визначається як

****

Вона є теоретичною оцінкою .

Медіана  є характеристикою, що вказує, де знаходиться «центр» розподілу:



*Означення 1*. Значення ознаки (варіанти), яка розділяє ранжований варіаційний ряд на дві рівні за кількістю елементів частини називається **медіаною вибірки** і позначається .

Якщо кількість елементів у вибірці є парною, тобто , тоді медіана визначається за таким співвідношенням

.

Якщо кількість елементів у вибірці є непарною, тобто , тоді медіана визначається за таким співвідношенням

.

Якщо дані представлені у вигляді групованої вибірки, то спочатку визначається медіанний інтервал  і використовується формула

,

де  - об’єм вибірки;

 - сума частот до медіанного інтервалу;

 - частота медіанного інтервалу.

Для того, щоб знайти медіанний інтервал, послідовно знаходять нагромаджені частоти . Першій нагромадженій частоті , яка більша за , відповідає медіанний інтервал (у випадку статистичного ряду  відповідає самій медіані).

Нагромаджена частота -го інтервалу визначається за формулою

.

**Приклад 1.**

Знайти медіану для наступної вибірки: 5, 3, 7, 10, 5, 5, 2, 10, 7, 2, 7, 7, 4, 2, 4.

*Розв’язання*.

Визначаємо об’єм вибірки, як загальну кількість елементів вибірки  З цього співвідношення знаходимо, що .

Будуємо варіаційний ряд вибірки

.

Знаходимо медіану

.

**Приклад 2.**

Знайти медіану для наступної вибірки: 5, 2, 10, 7, 2, 7, 7, 4, 2, 4.

*Розв’язання*.

Визначаємо об’єм вибірки, як загальну кількість елементів вибірки  З цього співвідношення знаходимо, що .

Будуємо варіаційний ряд вибірки

.

Знаходимо медіану

.

**Приклад 3.**

Задано груповану вибірку. Потрібно знайти її медіану.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Інтервал | [6,5; 6,9) | [6,9; 7,3) | [7,3; 7,7) | [7,7; 8,1) | [8,1; 8,5) | [8,5; 8,9) | [8,9; 9,3) |
| Частота | 3 | 10 | 20 | 32 | 16 | 12 | 7 |
| Нагромаджені  частоти | 3 | 13 | 33 | 65 | 81 | 93 | 100 |

*Розв’язання*.

Знаходимо об’єм вибірки .

Далі знаходимо . І заповнюємо останній рядок таблиці.

Бачимо, що для четвертого інтервалу нагромаджені частоти вперше перевищили . Це і буде медіанний інтервал. Тепер знаходимо саму медіану

.

**Приклад 4.**

Задано статистичний ряд вибірки. Потрібно знайти медіану вибірки.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | -2,8 | -1,6 | -0,4 | 0,8 | 2 |
|  | 3 | 2 | 7 | 2 | 6 |
| Нагромаджені  частоти | 3 | 5 | 12 | 14 | 20 |

Знаходимо об’єм вибірки .

Далі знаходимо . І заповнюємо останній рядок таблиці.

Бачимо, що для  нагромаджені частоти вперше перевищили . Це і буде медіана. Тобто, 

**Мода вибірки**

*Означення 2*. **Модою  вибірки**, яка відповідає дискретній випадковій величині, будемо називати елемент вибірки з найбільшою частотою.

Для групованої вибірки визначається модальний інтервал , якому відповідає найбільша щільність відносної частоти , де  - частота -го інтервалу,  - довжина -го інтервалу. Тоді мода **** шукається за формулою

,

де  - початок модального інтервалу;

 - кінець модального інтервалу;

 - частота інтервалу, що передує модальному;

 - частота модального інтервалу;

 - частота інтервалу наступного за модальним.

**Приклад 5.**

Знайти моду для наступної вибірки: 5, 2, 10, 7, 2, 7, 7, 4, 2, 4.

*Розв’язання*.

Будуємо варіаційний ряд вибірки

.

Будуємо статистичний ряд

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 2 | 4 | 5 | 7 | 10 |
|  | 3 | 2 | 1 | 3 | 1 |

Бачимо, що елементи 2 і 7 мають найбільші частоти, які співпадають між собою. Отже у нас буде дві моди  та .

**Приклад 6.**

Задано груповану вибірку. Потрібно знайти її моду.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Інтервал | [6,5; 6,9) | [6,9; 7,3) | [7,3; 7,7) | [7,7; 8,1) | [8,1; 8,5) | [8,5; 8,9) | [8,9; 9,3) |
| Частота | 3 | 10 | 20 | 32 | 16 | 12 | 7 |

*Розв’язання*.

До нашої початкової таблиці дописуємо ще два рядки

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Інтервал | [6,5; 6,9) | [6,9; 7,3) | [7,3; 7,7) | [7,7; 8,1) | [8,1; 8,5) | [8,5; 8,9) | [8,9; 9,3) |
| Частота | 3 | 10 | 20 | 32 | 16 | 12 | 7 |
|  | 0,4 | 0,4 | 0,4 | 0,4 | 0,4 | 0,4 | 0,4 |
|  | 7,5 | 25 | 50 | 80 | 40 | 30 | 17,5 |

Тепер визначаємо

; ; ; ; .

Далі шукаємо моду

.

**ЗАДАЧІ**

1) Нехай дані представлені у вигляді групованої вибірки

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Інтервал | [-6; -4) | [-4; -2) | [-2; 0) | [0; 2) | [2; 4) | [4; 6) |
| Частота | 7 | 4 | 9 | 5 | 5 | 10 |

Потрібно побудувати емпіричну функцію розподілу, гістограму та полігон частот.