Лабораторна робота №2

СТАТИСТИЧНИЙ АНАЛІЗ МАСИВІВ ДАНИХ

Мета роботи: навчитися створювати програми для дослідження статистичних властивостей масивів даних.

Теоретичні відомості

В теорії ймовірності використовуються різні поняття за допомогою яких можна описати властивості випадкових процесів. Математичним сподіванням, або середнім значенням дискретної випадкової величини є сума всіх її значень помножена на ймовірність їхньої появи:

$$M(x)=\sum_{i=1}^{N} x_i \cdot p_i \qquad (2.1)$$

Математичне сподівання також визначає середнє значення випадкової величини. Дисперсією випадкової величини називається математичне сподівання квадрату її відхилення відносно середнього значення. Дисперсія визначає розкид значень випадкової величини відносно математичного сподівання:

$$D(x)=M([x-M(x)]^2)=\sum_{i=1}^{N}(x_i-M(x))^2 \cdot p_i \qquad (2.2)$$

Середньоквадратичне відхилення випадкової величини визначається як корінь квадратний дисперсії:

$$\sigma=\sqrt{D(x)} \qquad (2.3)$$

Більше 90% всіх значень випадкової величини розподілених за нормальним законом розподілу знаходяться в інтервалі [-3σ, 3σ]. Якщо ймовірність появи випадкової величини є невідомою для визначення математичного сподівання використовується така формула:

$$M(x)=\frac{1}{N}\cdot\sum_{i=1}^{N} x_i \qquad (2.4)$$

де x_i — значення випадкової величини;

N — кількість спостережень.

Алгоритм знаходження математичного сподівання полягає у знаходженні суми всіх значень випадкової величини та її ділення на кількість спостережень.

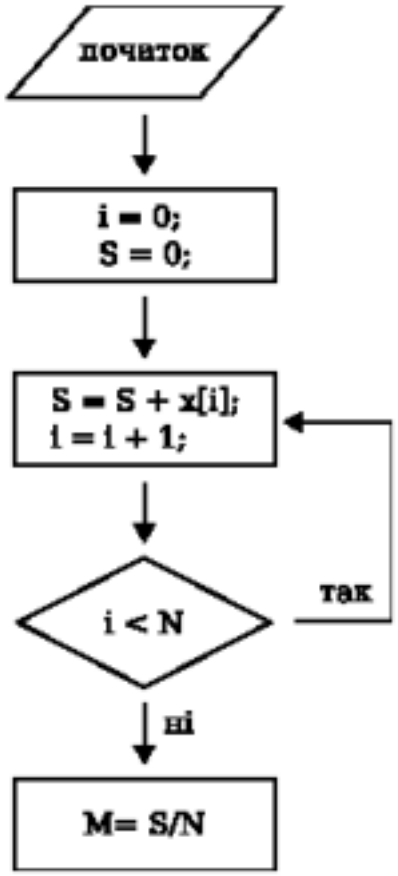


Рис. 2.1. Блок-схема алгоритму знаходження математичного сподівання випадкової величини

Дисперсія визначається за такою формулою:

$$D(x)=\frac{1}{N}\cdot\sum_{i=1}^{N}[M(x)-x_i]^2 \qquad (2.5)$$

Алгоритм знаходження дисперсії полягає у знаходженні суми квадратів відхилення всіх значень випадкової величини відносно математичного сподівання та її ділення на кількість спостережень.

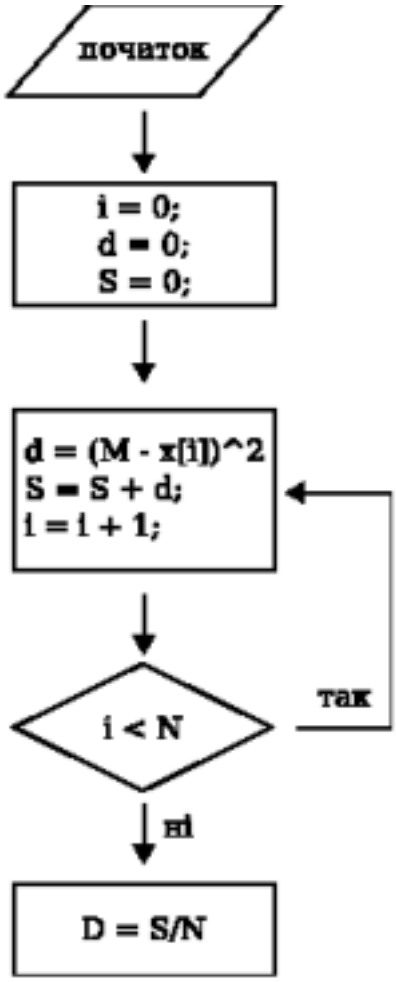
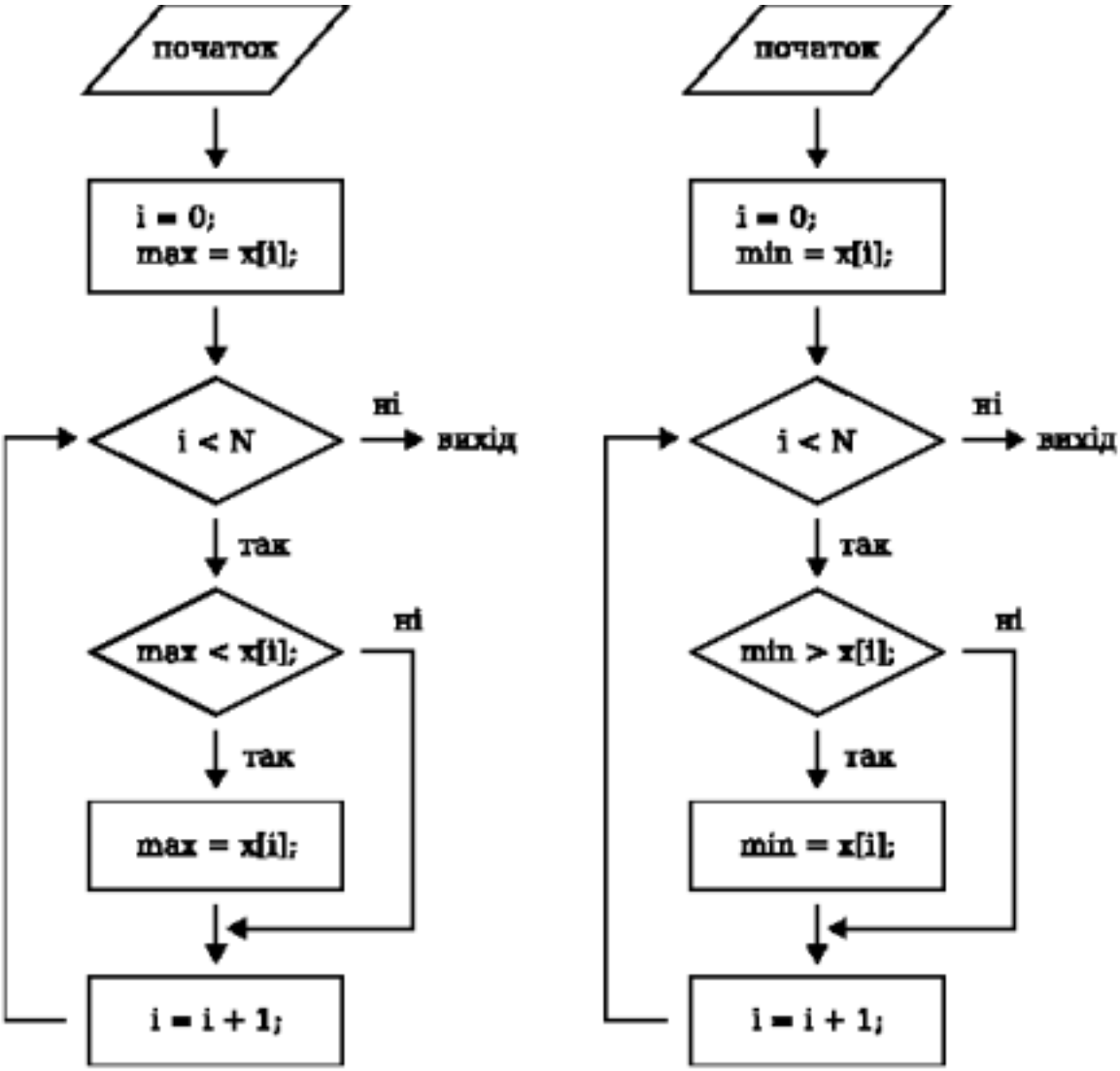


Рис. 2.2. Блок-схема алгоритму знаходження дисперсії випадкової величини

Математичне сподівання та дисперсія є основними статистичними характеристиками випадкової величини.

Для побудови гістограми необхідно розв'язати ряд задач серед яких найважливішими є знаходження ширини діапазону значень та визначення частоти потрапляння у інтервали. Задача пошуку даних полягає у знаходженні в послідовності елемента, або декількох елементів, із заданими властивостями його значення.

Прикладом такої задачі є пошук у масиві елементів з найбільшим, або найменшим значенням. Більш складною є задача пошуку кількох елементів із заданим значенням. Подібні задачі виникають дуже часто під час реалізації різних алгоритмів роботи програмного забезпечення.



а) б)

Рис. 2.3. Блок-схема алгоритмів пошуку:
а) максимального значення;
б) мінімального значення.

Пошук мінімального, або максимального значень здійснюється шляхом порівняння всіх елементів із еталонною змінною. Вона містить поточне значення випадкової величини, яке було визначене для порівняння, в процесі пошуку. На початку роботи алгоритму цій змінній присвоюється перше значення у масиві, а в процесі роботи — знайдене максимальне чи мінімальне значення. Після присвоєння еталонній змінній першого значення в масиві порівнюємо її з наступним елементом. Якщо виявиться, що вона є меншою (пошук максимального значення) чи більшою (пошук мінімального значення)

за досліджувану змінну, присвоюємо їй значення досліджуваного елемента. Далі переходимо до її порівняння з другим елементом досліджуваної послідовності і так далі. Порівняння необхідно провести для всіх елементів послідовності, щоб забезпечити правильність пошуку. Після його завершення еталонна змінна буде містити мінімальне, або максимальне значення із послідовності змінних (залежно від умови порівняння). В процесі пошуку для порівняння слід використати одну із логічних умов більше, або менше.

Для дослідження статистичних властивостей масивів даних використовується гістограма. На її основі можна встановити закон розподілу випадкової величини. Гістограма — це є спосіб представлення статистичних даних в графічному вигляді, наприклад у вигляді стовпців. Вона відтворює розподіл окремих вимірювань значень параметрів виробу, процесу чи явища. Іноді її називають частотним розподілом, оскільки гістограма показує частоту появи виміряних значень параметрів об'єкта.

Методика побудови гістограми

Гістограма будується у декілька етапів.

• Збір статистичних даних (формування вибірки). Під час цього етапу отримуються результати вимірювань параметра об'єкта, процесу чи явища. Для того, щоб за допомогою гістограми можна було встановити закон розподілу випадкової величини кількість проведених експериментів повинна бути не меншою тридцяти. Збільшення кількості експериментів на основі яких формується вибірка забезпечує підвищення точності встановлення статистичних властивостей. Для використання в програмі отриману вибірку значень випадкової величини можна зберегти за допомогою масиву.

• Визначення найбільшого та найменшого значення із отриманої вибірки. Використання цих показників забезпечує встановлення робочого діапазону в межах якого знаходяться всі значення.

• Робочий є необхідним для побудови гістограми і визначається як різниця між найбільшим та найменшим значеннями. На основі робочого діапазону можна визначити ширину інтервалів.

• Визначення необхідної кількості інтервалів в межах яких необхідно згрупувати результати вимірювань (встановити частоту появи значень в межах інтервалу). Кількість інтервалів може визначатися довільним чином, наприклад для забезпечення зручності відтворення гістограми на екрані монітору, або з

використанням логарифмічного масштабу. Чим більшою є кількість інтервалів гістограми тим точніше вона описує криву закону розподілу випадкової величини (за умови достатньої кількості спостережень).

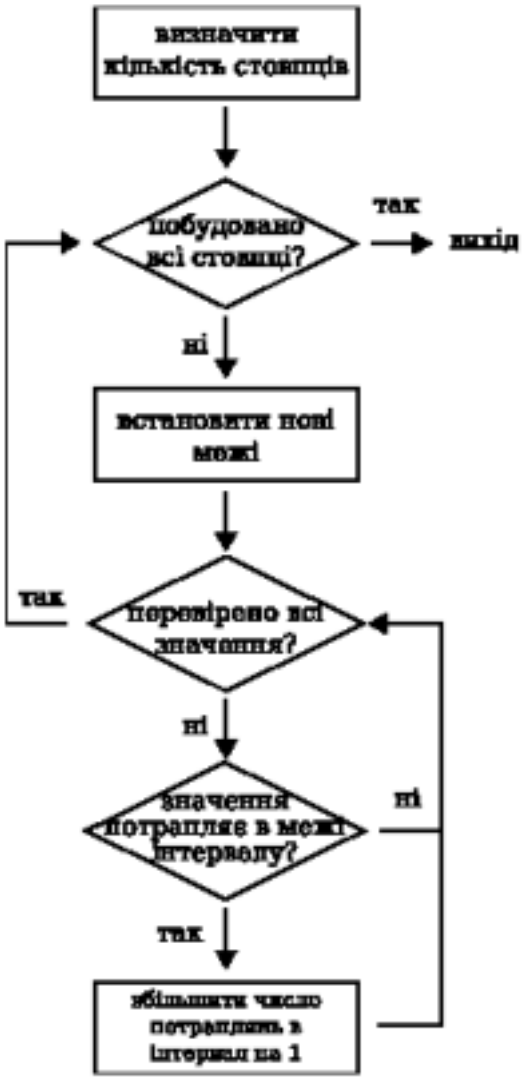


Рис. 2.4. Блок-схема алгоритму побудови гістограми

• Визначення меж інтервалів. Межі інтервалів необхідно встановити так, щоб отримані значення потрапляли у власні інтервали, ширина яких визначається на основі робочого діапазону значень та необхідної кількості стовпців на гістограмі. Можливі випадки, коли досліджувані значення потраплять на межу інтервалу і не будуть врахованими. Для уникнення цієї проблеми інтервали формують таким чином, щоб значення однієї із меж потрапляло в інтервал. Це означає, що необхідно вибрати правильну умову формування меж інтервалів. Прикладами такої умови є ( ], та [ ). Перша означає що права межа інтервалу

може дорівнювати значенню випадкової величини і їй відповідає логічний оператор <=. Друга — ліва межа інтервалу може дорівнювати значенню випадкової величини, а логічна умова >=.

• Визначення частоти потраплянь здійснюється шляхом підрахунку кількості випадкових величин, значення яких знаходяться у досліджуваному інтервалі. Чим більшою є частота потраплянь у інтервал, тим вищим є стовпчик гістограми. Підрахунок частоти появи значень можна здійснювати різними способами. Перший спосіб полягає у підрахунку кількості потраплять випадкової величини у заданий інтервал з переходом до наступного. Другий спосіб полягає у знаходженні інтервалу до якого належить значення випадкової величини з переходом до наступного її значення. Незалежно від того, який спосіб підрахунку частоти появи значень буде вибрано, програма буде виконувати подвійний цикл.

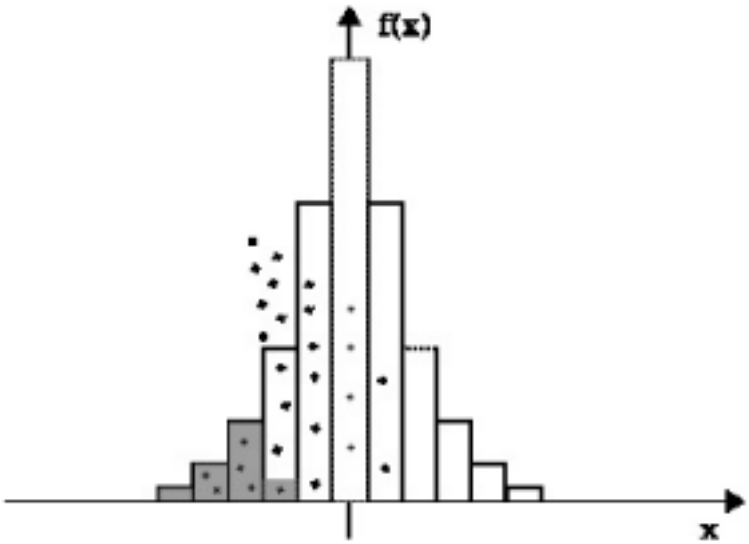


Рис. 2.5. Схема підрахунку частоти появи значень випадкової величини шляхом підрахунку кількості потраплять у заданий інтервал

• Після визначення робочого діапазону, ширини інтервалів та частоти появи значень переходять до етапу побудови самої гістограми. Для цього на осі абсцис позначають границі інтервалів, а по осі ординат — частоту появи значень, котрі потрапляють у конкретний інтервал. Висота стовпців є пропорційною частоті появи значень.

Для відтворення гістограми на екрані монітору в текстовому режимі необхідно використати символ заповнення 'x', 'o', тощо. Кількість символів у

рядку пропорційна висоті стовпця. На основі аналізу гістограми можна встановити закон розподілу випадкової величини.

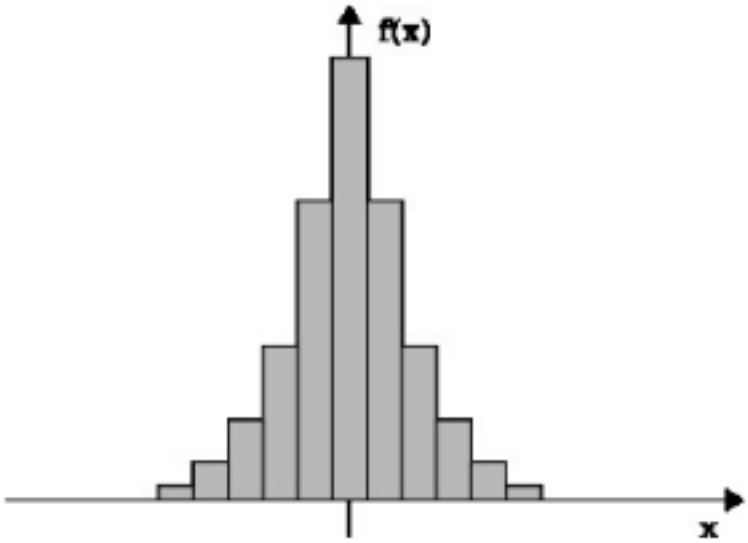


Рис. 2.6. Гістограма випадкової величини з нормальним законом розподілу

У випадку недостатньої кількості проведених спостережень, помилок одержання даних, або нестабільності досліджуваного процесу вигляд гістограми може відрізнятися від широко-відомих законів розподілу.

Порядок роботи

1. Запустити середовище розробки програм на мові C/C++ (BorlandC, GCC, MinGW, Dev-C++, Visual Studio, тощо).
2. Вибрати із залікової книжки дві останні цифри **m** (передостання) та **n** (остання). Якщо будь-яка із цифр рівна нулеві замінити її на 1.
3. Для виконання завдання попередньо заповнити масив g[] із 200 елементів випадковими значеннями з нормальним законом розподілу за допомогою удосконаленого перетворення Бокса-Мюллера.
4. Скласти програму для дослідження статистичних властивостей даних в масиві g[]. Визначити математичне сподівання, дисперсію та побудувати гістограму розподілу випадкової величини.
5. Додати до чисел в масиві значення **m**. Перемножити значення в масиві на **n**. Дослідити яким чином міняються статистичні параметри.

Контрольні запитання

1. Що таке випадкова величина?
2. Що таке закон розподілу випадкової величини?
3. Що таке математичне сподівання?
4. Що таке дисперсія?
5. Яким чином визначається математичне сподівання та дисперсія?
6. Як здійснюється пошук максимального та мінімального значень в масиві?
7. Властивості нормального закону розподілу?
8. Удосконалене перетворення Бокса-Мюллера?
9. Що таке гістограма?
10. Яким чином побудувати гістограму?

Зміст звіту

1. Титульний лист
2. Тема та мета роботи
3. Короткі теоретичні відомості
4. **Результати виконаної роботи**
5. **Висновок**

ЛІТЕРАТУРА

1. Березин Б.И., Березин С.Б. Программирование на С и С++ - М.: ДИАЛОГ-МИФИ, 2001. - 288 с.
2. Керниган Б., Ритчи Д.. Язык программирования С, 2-е издание - М.: Вильямс — 2009. - 292 с.
3. Кибзун А.И. Теория вероятности и математическая статистика. Базовый курс с примерами и задачами / Учебн. Пособие. - М.:ФИЗМАТЛИТ, 2002. -224 с.
4. Лафоре Л. Объектно-ориентирование программирование в С++, 4-е издание — М.: Питер, 2004. - 923 с.
5. Минашкин В.Г. Теория статистики: учебно-методический комплекс. - М.: Изд. Центр ЕАОИ. 2008. - 296 с.
6. Папас К., Мюррей У. Программирование на С и С++ - К.: Издательская група BHV, 2000. - 320 с.
7. Пугачев В.С. Теория вероятности и математическая статистика: Учеб. Пособие. - 2-е изд., исправл. и дополн. - М.:ФИЗМАТЛИТ, 2002. - 496 с.
8. Шилдт Г. Справочник программиста С/С++, 3-е изд.: Пер. с англ. - М. Издательский дом “Вильямс”, 2003. - 432 с.