1 ОПИСОВІ СТАТИСТИКИ

1.1 Мета роботи

Отримання студентами практичних навичок розрахунку описових статистик: математичного сподівання, дисперсії, середньоквадратичного відхилення.

1.2 Теоретичні відомості

Описові статистики служать для опису найзагальніших характеристик спостережуваних величин.

До числа основних описових статистик відносяться: середня, вибіркова дисперсія, стандартне відхилення, медіана, мода, максимальне і мінімальне значення, розмах, квантилі [1, с.38].

Середнє М, або, точніше, оцінка середнього, обчислюється просто як середнє арифметичне спостережень. Оцінку середнього називають також вибірковим середнім. Нехай ви спостерігаєте значення X(1)...X(N), наприклад, відзначаєте час, коли ви прокидаєтеся вранці.

Формула для вибіркового середнього має наступний вигляд:

|  |  |
| --- | --- |
| M = (X (1 )+... X (N)) / N | (1.1) |

Якщо протягом трьох днів ви прокидалися о 7.00, 8.30, 6.15, то середній час вашого підйому – 7год. 15 хв.

Вибіркове середнє є тією точкою, сума відхилень від якої усіх розглянутих спостережень дорівнює 0.

Формально це записується наступним чином:

|  |  |
| --- | --- |
| (M-X (1)) + (M-X (2 ))+...+( M-X (N)) = 0 | (1.2) |

При розрахунку вибіркового середнього необхідно переконатися що дана властивість дійсно має місце, тобто сума відхилень спостережуваних значень від середнього арифметичного дійсно дорівнює 0, використовуючи формули 1.1 та 1.2.

Вибіркове середнє – єдина точка, яка має таку властивість, і це виділяє її серед всіх інших.

Крім того, вибіркове середнє має ще одну чудову властивість: сума квадратів відстаней між значеннями, що спостерігаються, та їх середнім арифметичним є мінімальною. Якщо замість середнього арифметичного взяти будь-яку іншу величину, то сума квадратів відстаней, що спостерігаються, і цією величиною буде тільки більше, але ніяк не менше [2, с.182].

Наприклад, у таблиці 1.1 наведені значення доходу фірми, виражені в мільйонах, та середній дохід у день. Доходи вимірювалися протягом 5 днів (робочий тиждень).

Таблиця 1.1 – Доходи фірми

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| понеділок | вівторок | середа | четвер | п’ятниця | середній дохід у день |
| 100 | 102 | 81 | 105 | 96 | 96,80 |

Іншими важливими статистичними показниками є вибіркова дисперсія та стандартне відхилення [3, с. 90].

Вибіркова дисперсія визначається я за формулою 1.3.

|  |  |
| --- | --- |
| ((X (1)-m (n )))2 +...+( X (N)-m (n))2)/(n-1) | (1.3) |

Вибіркова оцінка є переконлива оцінка дисперсії розподілу.

Корінь квадратний з вибіркової дисперсії є стандартним відхиленням.

Стандартне (середнє квадратичне) відхилення – ступінь відхилення усіх значень ознаки від свого середнього показника – є одним із найважливіших методів, що допомагають визначити, наскільки змінюється певна величина: чим більше стандартне відхилення, тим ширший діапазон змін значень цієї величини.

Стандартне відхилення дорівнює квадратному кореню із суми квадратів відхилень спостережуваних значень від середнього, поділеній на (N – 1), де N – число спостережень.

Стандартне відхилення є важливою характеристикою вибірки даних.

До інших важливих характеристик відносяться: мода, медіана, квантиль.

Мода – це значення розподілу (відповідно до звичайного значення цього слова), що найбільш часто зустрічається. Вона добре описує, наприклад, типову реакцію водіїв на сигнал світлофора про припинення руху.

Розподіл з одним максимумом щільності розподілу називається «унімодальним», з двома максимумами – «бімодальним». Якщо багатокутник розподілу або крива розподілу мають більше одного максимуму, розподіл називають «полімодальним». Якщо розподіл має посередині НЕ максимум, а мінімум, то він називається «антимодальним».

Наявність більш ніж однієї моди, часто вказує на різнорідність статистичного матеріалу, який був покладений в основу дослідження.

Мода і середня величина по-різному характеризують сукупність. Мода визначає безпосередньо розмір ознаки, притаманної хоча і значній частині сукупності, але ж не всій сукупності. Мода за своїм узагальнюючим значенням менш точна порівняно з середньою арифметичною, яка характеризує сукупність в цілому з урахуванням усіх без винятку елементів сукупності.

Мода використовується коли нам треба швидка і приблизна міра центральної тенденції та коли потрібна міра центральної тенденції, що має бути типовим значенням.

Переваги моди:

* мода показує найбільш поширене значення в розподілі;
* на моду не впливають екстремальні значення – так як на середнє;
* моду можна визначити для відкритих інтервалів/категорій;
* допомагає аналізувати якісні дані;
* моду можна виявити просто побудувавши графік розподілу чи стовпчасту діаграму.

Обмеження:

* не включає до визначення/розрахунку всі спостереження розподілу, а лише концентрацію частот;
* подальші алгебраїчні перетворення неможливі – на відміну від середнього;
* буває важко визначити моду у випадку полімодального чи бімодального розподілу.

Медіана – це середина спостереження у вибірці.

Медіаною називають таке значення ознаки, яке поділяє ранжируваний ряд розподілу на дві рівні частини, тобто значення, яке перебуває в середині ряду розподілу. Якщо в дискретному варіаційному ряду 2n + 1 випадків, то значення ознаки у випадку n + 1 є медіанним.

Більш точно: нехай є вихідна вибірка даних, що мають чисельне значення X(1)… X(2)… X(n).

Впорядкуємо їх за зростанням. Впорядкування за зростанням значень називається варіаційним рядом (формула 1.4).

|  |  |
| --- | --- |
| X(1)< X(2)<…< X(n) | (1.4) |

Серединне значення в цьому ряді називається медіаною [4, с.59].

Якщо в ряду парне число 2n випадків, медіану визначають як середню арифметичну з двох серединних значень. Наприклад, якщо 15 студентів групи розташувати у порядку зростання середнього балу, тобто в ранжируваний ряд за середнім балом, то середній бал у восьмого студента буде медіанним. Якщо ж число студентів буде 16 чоловік, то медіаною буде середнє значення балів зерна восьмого і дев'ятого студентів.

Якщо бути більш точним, то слід сказати так: при непарній N медіана є значенням із ряду з номером (N – 1)/2, при парному N медіана є напівсумі значень з номерами (N/2) і (N/2 + 1).

Медіана є кращою характеристикою центральної тенденції, коли межі крайніх інтервалів відкриті. Медіана є найприйнятнішою характеристикою і в разі, якщо в ряду розподілу є значно великі і значно малі значення, які впливають на середню величину, а на медіану – ні. Медіана, крім того, володіє властивістю лінійного мінімуму: сума абсолютних значень відхилень величини ознаки у всіх одиниць сукупності від медіани – найменша.

Ця властивість має велике значення для вирішення деяких практичних завдань – наприклад, для розрахунку самої короткої з усіх можливих відстаней для різних видів транспорту, для розміщення станцій технічного обслуговування таким чином, щоб відстань до всіх машин, що обслуговуються даною станцією, була мінімальною і т. п.

Для чого використовують медіану?

* коли потрібно знайти точну середню точку, точку на «півдорозі» від найменшого значення до найбільшого;
* коли екстремальні значення впливають на середнє – медіана є найкращою мірою центральної тенденції;
* медіану використовують коли потрібно, щоб певні значення впливали на центральну тенденцію, але все, що про них відомо – що вони «нижче» або «вище» медіани.

Переваги медіани:

* легко вирахувати та зрозуміти;
* для підрахунку медіани не потрібні всі значення в розподілі;
* екстремальні значення розподілу не впливають на медіану;
* її можна визначити і для «відкритих» категорій/класів інтервалів.

Обмеження медіани:

* вона не так жорстко визначена як середнє, оскільки її значення не так вираховується, як знаходиться (серед значень в розподілі);
* не враховує всі спостереження (значення для всіх спостережень);
* з медіаною потім не можна робити алгебраїчні перетворення так, як із середнім;
* потребує впорядкування значень або класів інтервалів у висхідному чи спадному порядку;
* часом медіаною може бути значення, не присутнє у самому розподілі.

В відсортованій вибірці X(1) – мінімальне значення, X(n) – максимальне значення. Різниця між максимальним і мінімальним значенням називається розмахом.

Квантилі відсікають в межах ряду певну частину його членів. Тобто, квантиль (термін використаний вперше Кендалом в 1940 р.) розподілення значень – це таке число xp, що значення p-ї частини сукупності менше або рівне xp. Наприклад, квантиль 0,25 (також називається 25-процентилем або нижнім квартилем) змінної – це таке значення xp, що 25% (p) значень змінної попадають нижче даного значення

1.3 Завдання до виконання роботи

Реалізуйте за допомогою обраної мови програмування обчислення таких значень описової статистики:

* + вибіркове середнє значення;
  + вибіркова дисперсія;
  + стандартне відхилення;
  + медіана;
  + мода;
  + максимальне і мінімальне значення;
  + розмах;
  + квантилі: 0.1, 0.25, 0.5, 0.75
  1. Ускладнене завдання із використанням набору даних

Завантажте датасет доступний за адресою:

**<https://www.kaggle.com/tmdb/tmdb-movie-metadata>**

Одним з етапів дослідження даних є «Попередній аналіз». Для виконання попереднього аналізу даних застосовують засоби та методи, що вивчаються в курсі «Теорія ймовірностей та математична статистика».

В ході попереднього аналізу відповідають на ряд запитань, частина яких безпосередньо пов’язана з темою даної лабораторної роботи. Наприклад, середні значення, мінімальне і максимальне значення в наборах даних і т.д.

Відповідно, для зазначеного вище набору даних TMDB необхідно:

1. Для числових наборів даних створити гістограми та візуалізувати їх
2. Для нечислових наборів підрахувати частоти окремих значень і візуалізувати.
3. Розрахувати метрику популярності фільмів беручи до уваги кількість переглядів фільму. Тобто, середнього значення рейтингу недостатньо, наприклад фільм може бути переглянутий 10000 разів і отримати ретинг 5.6, а може бути переглянутий тричі і мати рейтинг 9 (формула 1.5).

|  |  |
| --- | --- |
|  | (1.5) |

де v – кількість оцінок за фільм;

m – мінімальна кількість оцінок у вибірці;

R – середній рейтинг конкретного фільму;

С – середнє значення рейтингу в датасеті;

1. На основі отриманого рейтингу сформуйте Top-10 фільмів у тренді для цього набору даних.
2. Сформуйте Top-5 фільмів для різних жанрів, що наявні в таблиці
3. Замініть С на медіану, і перерахуйте Top-5 жанрів і Top-10.

1.5 Зміст звіту

Звіт повинен містити наступні частини:

* + тема;
  + мета;
  + масив значень вхідної вибірки;
  + перелік описових статистик, що обчислюються;
  + частини програмного коду, що реалізують визначену величину описової статистики з поясненням;
  + скріншот результатів виконання програми;
  + висновки.

1.6 Висновки

У даній лабораторній роботі було:

* пояснено призначення базових статистичних величин: математичного сподівання, дисперсії, середньоквадратичного відхилення;
* наведено приклади розрахунку базових статистичних величин;
* запропоновано самостійні завдання, які дозволять закріпити теоретичний матеріал щодо розрахунку базових статистичних величин.

1.7 Контрольні запитання

1. Що таке математичне сподівання? Наведіть формулу.
2. Наведіть приклад використання математичного сподівання.
3. Що таке дисперсія? Наведіть формули.
4. Наведіть приклад використання дисперсії.
5. Що таке середньоквадратичне відхилення? Наведіть формулу.
6. Коли використовується мода?
7. Переваги та обмеження моди.
8. Переваги та обмеження медіани.
9. Що таке квантиль? Які спеціалізовані види квантилів існують?
10. В яких статистичних розрахунках використовуються квантилі?