

Під час аналізу даних виділяються наступні етапи: отримання вхідної інформації, безпосередньо сама обробка її, аналіз та інтерпретація результатів обробки даних.

Головне зробити правильні висновки з результатів.

Значення змінних які спостерігаються можуть бути як *кількісні* так і *якісні*. Якісні змінні поділяють на *ординальні* та *номінальні*. Ординальні змінні називають *порядковими*, а номінальні – *класифікаційними*. Обидва типи змінних приймають свої значення з деякої множини, елементи якої називають *градаціями*. Градації, які приймає як свої значення ординальна змінна, природно **впорядковані за ступенем прояву властивості**. Градації номінальної змінної такого порядку **не мають**. Серед якісних змінних виділяють *категоризовані* та *не категоризовані*.

До категоризованих змінних відносять змінні, для яких повністю визначена множина градацій та правило віднесення значення змінної, яке спостерігається, до певної градації.

Змінні ще поділяють на *дискретні* та *неперервні*.

1 Групування змінних

ξ – скалярна змінна, яка досліджується.

Вибірка об'єму n : x_1, x_2, \dots, x_n .

У випадку великих об'ємів вибірок виникає бажання провести деяке перетворення їх з метою стиснення даних без суттєвої втрати вибірками інформативності, а тільки згодом проводити обробку цих перетворених даних. Як правило, його застосовують при обробці спостережень над неперервними змінними, коли об'єм вибірки перевищує 50, а над дискретними змінними, коли кількість значень m , які вони приймають, перевищує 10.

Перехід до згрупованих даних:

1. Визначити $x_{\min} = \min_i(x_i)$, $x_{\max} = \max_i(x_i)$;
2. Інтервал $[x_{\min}, x_{\max}]$ розбивають на s однакових під-інтервалів $[a_i, b_i)$, $i = \overline{1, s}$. Зазвичай $5 \leq s \leq 30$. Зазвичай $s = 1 + \lceil \log_2 n \rceil$ або $s = \lceil 10 \log_{10}(n) \rceil$;
3. $x_i^* = \frac{a_i + b_i}{2}$ – центральна точка.

v_i – кількість вимірів з вибірки що належать інтервалу $[a_i, b_i)$.

$$\{x_1, x_2, \dots, x_n\} \mapsto \{x_i^*, v_i\}_{i=1}^s \left(\sum_{i=1}^s v_i = n \right).$$

Рекомендується $v_i \geq 5$, в разі $v_i < 5$ сусідні інтервали зливаються в один.

Зауваження! При проведенні групування даних зовсім не обов'язково брати під-інтервали однакової довжини.

$F_\xi(x) = P\{\xi < x\}$ – функція розподілу, $p_\xi(x)$ – функція щільності, $\{y_i, p_i\}_{i=1}^m$ – полігон ймовірності, якщо ξ – дискретна випадкова величина, що набуває значення y_i з ймовірністю p_i , $i = \overline{1, m}$.

Оцінка характеристик по згрупованим даним:

Емпірична (вибіркова) функція розподілу $\hat{F}_\xi(x)$ буде $\hat{F}_\xi(x) = \frac{1}{n} \sum_{i: b_i \leq x} v_i$.

Емпірична (вибіркова) функція щільності $\hat{p}_\xi(x)$ буде $\hat{p}_\xi(x) = \frac{v_{i(x)}}{n(b_{i(x)} - a_{i(x)})}$, де $i(x)$ – номер підінтервалу якому належить x .

2 Моделювання змінних

Потреба в генерації спостережень над випадковими величинами із заданими функціями розподілу.

Зазвичай $\xi = g(\xi_1, \xi_2, \dots, \xi_q)$, де $\xi_1, \xi_2, \dots, \xi_q$ – найпростіші випадкові величини, як правило вони рівномірно розподілені на відрізку $[0, 1)$.

Датчик (генератор) випадкових чисел – спеціальний пристрій, який після запиту на виході дозволяє отримати реалізацію випадкової величини із заданим законом розподілу.

Класи датчиків (генераторів) випадкових чисел:

- **табличні** – таблиця, заповнена реалізаціями випадковою величини із заданим законом розподілу, зазвичай досить високої якості, але вони маю обмежений об'єм. Кількість вибірок невелика.
- **фізичні** – деякий електронний пристрій на виході якого отримують необхідну реалізацію вибірки довільного об'єму, але кожна вибірка унікальна і неповторна.
- **програмні** – програма, що формує потрібну реалізацію. Базуються на використанні рекурентних формул з деякою глибиною пам'яті: задаючи однакові початкові значення можна отримати однакові вибірки. Генератор періодичний, отримані числа