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 степенем 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: $x_1, x_2, \ldots, x_n$.

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 $x_{\min} = \min\limits_i(x_i)$, $x_{\max} = \max\limits_i(x_i)$;
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 $s = 1 + [\log_2 n]$ або $s = [10 \log_{10}(n)]$;
3. $x_i^* = \dfrac{a_i + b_I}{2}$ – центральна точка.

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

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

   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 $\hat{F}_\xi(x) = \frac{1}{n} \sum\limits_{i:b_i \le x} v_i$.
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 "псевдовипадкові".

3 Програмні датчики

Генератор випадкової величини з $F(x) = U([0, 1))$.

Лінійна змішана формула:

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M} \\ \tilde{x}_i = \left(a_0 + \sum\limits_{j=1}^{\ell} a_j \tilde{x}_{i-j} \right) \mod M, i = 1, 2, \dots \end{cases}$$

$\ell \ge 1$, $a_j \ge 0$ $(j = \overline{1,\ell})$, $M > 0$, ℓ, a_j $(j = \overline{0,\ell})$, $M \in \mathbb{Z}^+$, $0 \le \tilde{x}_{\ell-j} \le M - 1$, $j = \overline{1,\ell}$.

Мультиплікативний конгруентний метод: Лінійна змішана формула ($\ell = 1$, $a_0 = 0$).

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M} \\ \tilde{x}_i = (a_1 \tilde{x}_{i-1}) \mod M, i = 1, 2, \dots \end{cases}$$

$0 \leq \tilde{x}_0 \leq M-1$, $\{\tilde{x}_i\}_{i\geq 0} \in \{0, 1, \ldots, M-1\}$.

Послідовність $\{\tilde{x}_i\}_{i\geq 0}$ періодична. $T_{\max}$ – максимальний період. $T_{\max} \leq M$. Вигідно взяти M якомога більшим, ближчим до максимального цілого числа, наприклад найбільше просте число, що менше $\max \text{int}$.

Мультиплікативний конгруентний метод не дозволяє досягти максимального теоретично можливого періоду рівного M.

$$\lambda(M) = \begin{cases} 1, & M = 2 \\ 2, & M = 4 \\ p^{q-1}(p-1), & M = p^q, p > 2, p \in \mathbb{P}, q \geq 1 \\ \text{lcm}\left(\lambda\left(p_1^{q_1}\right), \lambda(p_2^{q_2}), \ldots, \lambda(p_k^{q_k})\right), & M = p_1^{q_1} \cdot p_2^{q_2} \cdot \ldots \cdot p_k^{q_k}. \end{cases}$$

Теорема 3.1. Максимальний період послідовність $\{\tilde{x}_i\}_{i\geq 0}$ мультиплікативного конгруентного методу $T_{\max} = \lambda(M)$. $T_{\max}$ досягається при:

1. $\gcd(\tilde{x}_0, M) = 1$;
2. $a_1^{\lambda(M)} \bmod M = 1$, a_1 є первісним коренем за модулем M.

Зауваження. Якщо покласти M рівним простому числу, то $T_{\max} = M - 1$. В залежності від розрядності комп'ютера найбільшим простим числом буде:

розрядність	16	32	64
max просте число	$2^{16} - 15$	$2^{32} - 5$	$2^64 - 59$

Змішаний конгруентний метод:

Лінійна змішана формула ($\ell = 1$, $a_0 > 0$).

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M} \\ \tilde{x}_i = (a_0 + a_1\tilde{x}_{i-1}) \bmod M, i = 1, 2, \ldots \end{cases}$$

Теорема 3.2. Для отримання послідовності $\{\tilde{x}_i\}_{i\geq 0}$ яка досягає свого max періоду $T_{\max} = M$, необхідно:

- $\gcd(a_0, M) = 1$;
- $(a_1 - 1) \bmod p = 0$ для всіх $p|M$, $p \in \mathbb{P}$;
- $(a_1 - 1) \bmod 4 = 0$, якщо $4|M$.

Зауваження! Вибір параметрів змішаного конгруентного методу не є гарантією високої якості вибірки. Наприклад $a_0 = a_1 = 1$.

Квадратичний конгруентний метод:

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M} \\ \tilde{x}_i = (a_0 + a_1\tilde{x}_{i-1} + a_2\tilde{x}_{i-1}^2) \bmod M, i = 1, 2, \ldots \end{cases}$$

$T_{\max} = M$.

Ускладнення лінійної змішаної формули:

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M} \\ \tilde{x}_i = g(\tilde{x}_{i-1}, \tilde{x}_{i-2}, \ldots, \tilde{x}_{i-\ell}) \bmod M, i = 1, 2, \ldots \end{cases}$$