Числові характеристики випадкових величин, показники варіації; первинна статистична обробка кількісних ознак

1. Генеральна сукупність та вибірка. Репрезентативність вибірки
2. Параметри генеральної сукупності і вибіркові характеристики
3. Оцінки генеральних параметрів за вибірковими характеристиками
4. Міри положення, міри розсіювання і міри форми при характеризуванні вибірки
5. Довірчий інтервал для середнього арифметичного

1. Генеральна сукупність та вибірка. Репрезентативність вибірки

- **Генеральна сукупність (N)** – сукупність, з якої обирають певну її частину для сумісного дослідження

↓

- **Вибіркова сукупність** (**вибірка**) **(n)**

- Формування вибірки – ***повторна*** і ***безповторна вибірки***
- ***Репрезенативність вибірки*** – формування вибірки, коли вона найбільш повно представляє властивості генеральної сукупності
- Метод досягнення – ***рандомізація*** – відбір об’єктів у вибірку з генеральної сукупності випадковим чином.

2. Параметри генеральної сукупності і вибіркові характеристики

- Генеральна сукупність характеризується – **генеральними параметрами**

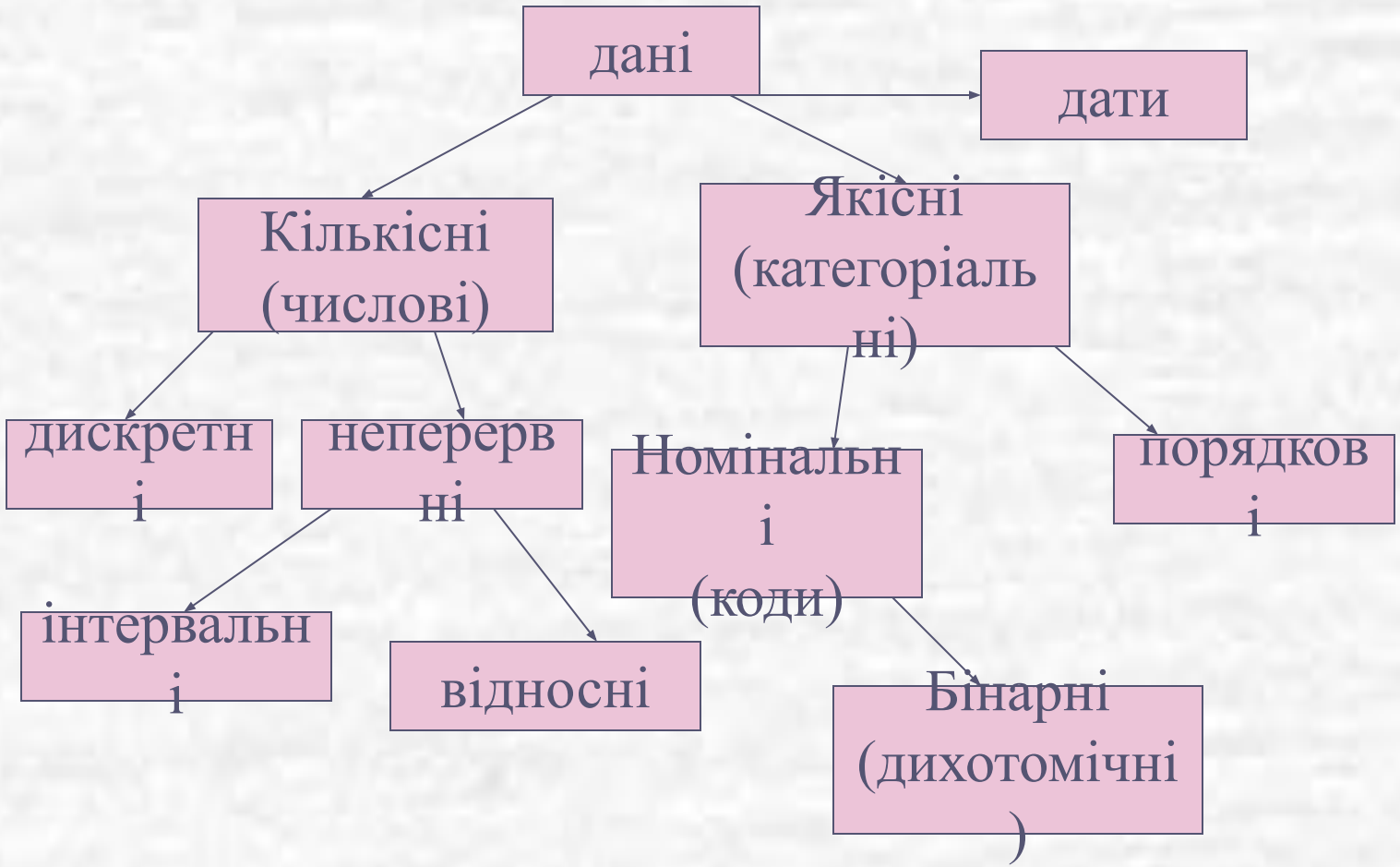
- Вибірка характеризується – **вибірковими характеристиками**, які наближаються до генеральних параметрів, але не дорівнюють їм

Незміщені, ефективні

Статистичні похибки – вказують на величину відхилення вибіркової характеристики від відповідного генерального параметра

- **Точкові характеристики** (міри положення, міри розсіювання, міри форми)
- **Інтервальні характеристики** (довірчий інтервал для середнього)

Класифікація даних

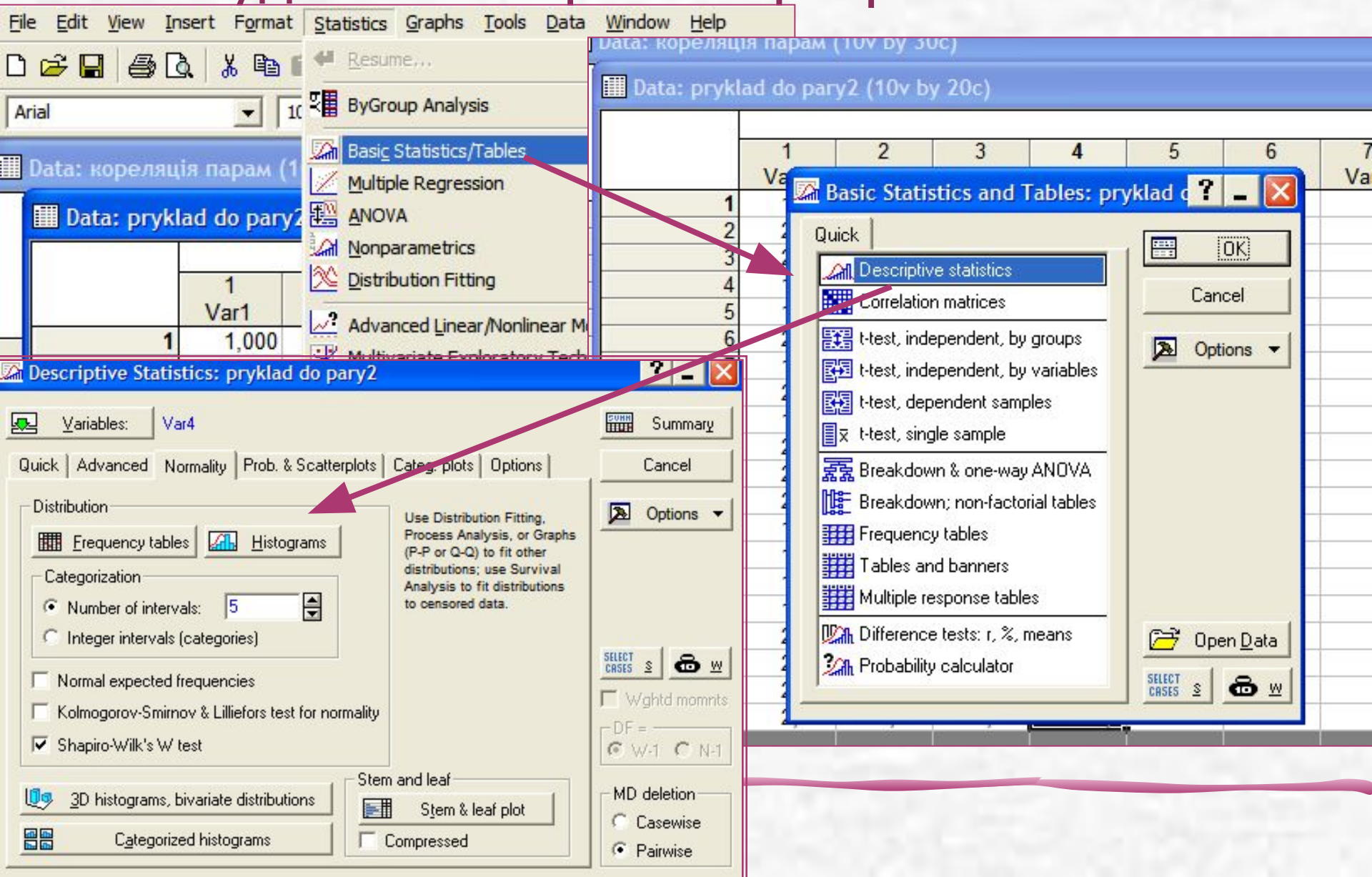


Попереднє впорядкування даних

- **Ранжування** – розміщення всіх значень ознаки x_i в порядку зростання (спадання)
- **Ряд розподілу** – ряд ранжованих даних, в якому розмах варіації (x_{min} – x_{max}) розбивають на рівні інтервали (**класи**) і шукають частоту зустрічаємості значень в кожному класі

- **Гістограма** – графік розподілу частот

Побудова гістограм в програмі Statistica



Розбиття вибірок на класи

- Правило Старджеса:

Число класів – k:

k = 1 + 3.31*lg (n)

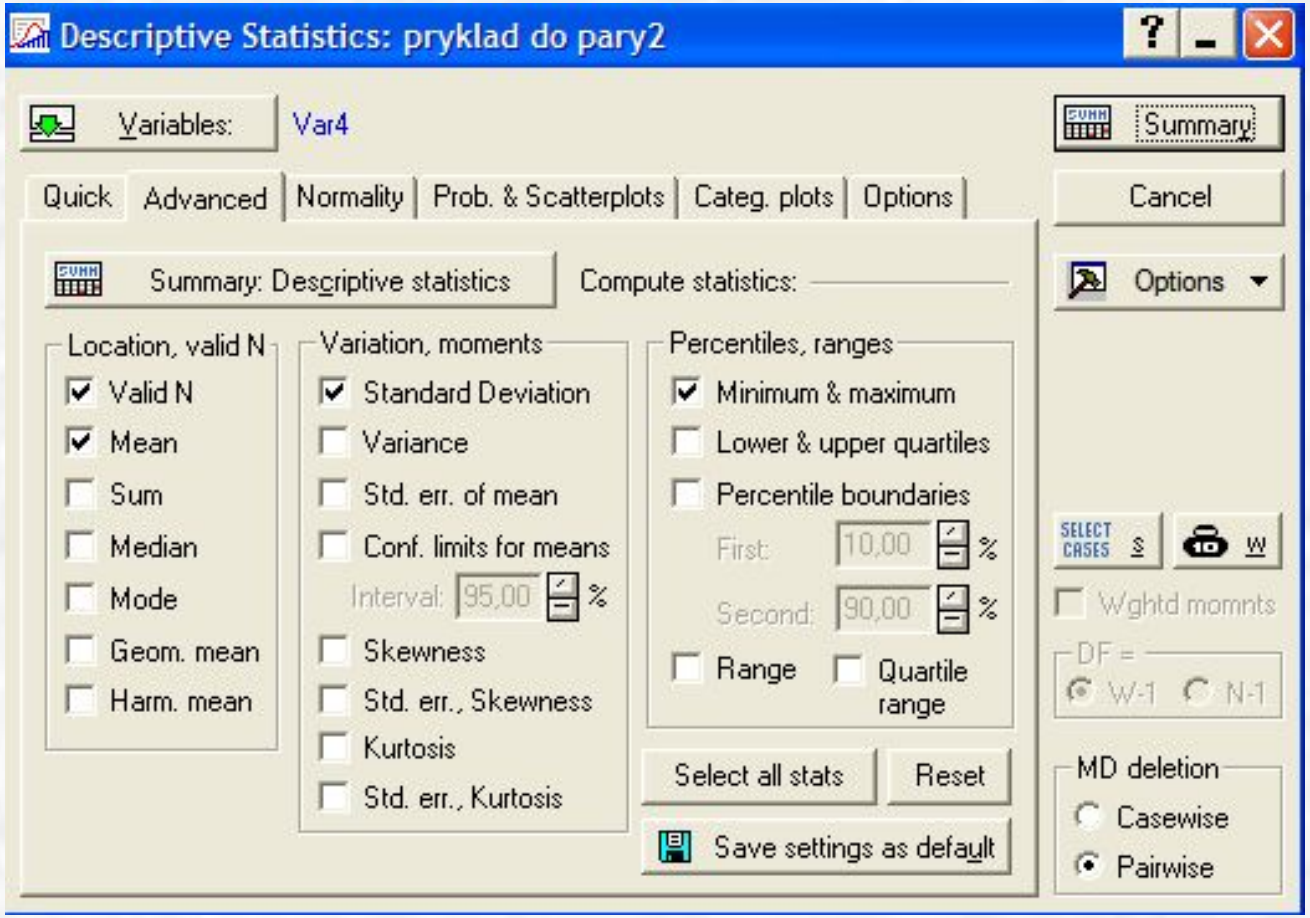
n	10-20	30-50	60-90	100-200	300-400	500-800	900-1500	2000
k	4	5-6	7	8	9	10	11	12

Приклад:

- Дані по захворюваності на грип у районній поліклініці згрупували за віком. Знайти міри положення цієї вибірки:

вік	20-29	30-39	40-49	50-59	60-69
Кількість хворих	45	36	175	361	825
Накопичені частоти	45	81	256	617	1442

Вибіркові характеристики:



3. Міри положення

Середнє арифметичне (mean)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{j=1}^{k} x_j n_j}{\sum_{j=1}^{k} n_j}$$

- X_i - значення (точка) вибірки,
- n – загальний об'єм вибірки

- X_j – значення вибірки коли воно зустрічається декілька разів (серединне значення інтервалу),
- n_j – частота, з якою спостерігається значення x_j (об' єм інтервалу)
- k – кількість інтервалів

$$\bar{x} = [24.5 * 45 + 34.5 * 36 +$$
$$+ 44.5 * 175 + 54.5 * 361 +$$
$$+ 64.5 * 825] / 1442 = 58.57$$

Медіана (median)

$$Me = x_{Me} + \frac{h(\frac{\sum m_x}{2} - m_x^{\max})}{m_m}$$

- це значення, яке ділить ранжований варіаційний ряд на 2 рівні за об'ємом групи

m_x – середина вибірки (1/2 вибірки)

h – ширина інтервалу,

m_m – об'єм медіанного інтервалу,

x_{Me} – початок медіанного інтервалу,

m_x^{max} – частота, накопичена на початок медіанного класу

$$Me = 60 + \frac{10 * 721 - 617}{825} = 61.26$$

Мода (mode)

$$M_o = x_{Mo} + \frac{h(m_{Mo} - m_{Mo-1})}{2m_{Mo} - m_{Mo-1} - m_{Mo+1}}$$

- це значення, яке спостерігається найбільшу кількість разів

- x_{Mo} – початок модального інтервалу,
- h – ширина інтервалу,
- m_{Mo} – об'єм модального інтервалу,
- m_{Mo-1} – об'єм інтервалу перед модальним
- m_{Mo+1} – об'єм інтервалу після модального

$$Mo = 60 + \frac{10*(825-361)}{2*825-361-0} = 65$$

Міри розсіяння (варіації)

- показують розкид даних у вибірці відносно середнього значення

- **Варіаційний розмах (розмах, range)**

$$R_v = x_{\max} - x_{\min}$$

$$R_v = 69 - 20 = 49$$

- **Емпірична дисперсія (вибіркова дисперсія) (variance)**

$$D = s^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}$$

$$D = \frac{[(24.5 - 58.57)^2 + (34.5 - 58.57)^2 + (44.5 - 58.57)^2 + (54.5 - 58.57)^2 + (64.5 - 58.57)^2]}{1442 - 1} =$$

$$= 1.38$$

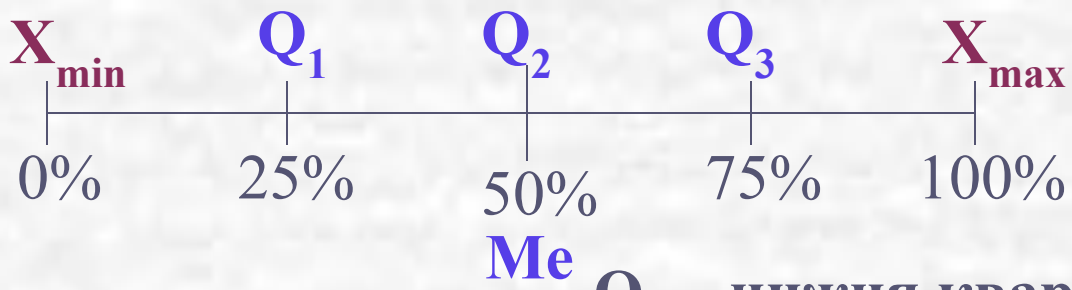
Стандартне відхилення (середнє квадратичне відхилення) (standard deviation)

$$\sigma = s = \sqrt{D}$$

$$\sigma = \sqrt{1.38} = 1.175$$

Інтерквартильний розмах (quartile range)

$$Q_3 - Q_1$$



Q_1 – нижня квартиль (lower quartile)

Q_3 – верхня квартиль (upper quartile)

Перцентіль – значення, яке міститься на межі певного % ранжованої вибірки

Міри форми

- Асиметрія (skewness) – вказує, наскільки розподіл симетричний відносно середнього (позитивна і негативна асиметрія)

$$\frac{n}{(n-1)(n-2)} * \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^3$$

- Ексцес (kurtosis) – міра гостроверхості відносно нормального розподілу (позитивний і негативний)

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} * \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Довірчий інтервал для генерального середнього

- ***Довірчий інтервал*** – інтервал, відносно якого з початково заданою ймовірністю P (P=1-α) можна стверджувати, що він містить невідоме значення генерального параметра
- P – ***довірча ймовірність***,
- α – ***рівень значущості***
- ***Довірчий інтервал для генерального серед limits of mean)***:

P = 0.95	**t1 = 1.96**
P = 0.99	**t2 = 2.58**
P = 0.999	**t3 = 3.29**

$$\bar{x} - t\frac{\sigma}{\sqrt{n}} \le M(x) \le \bar{x} + t\frac{\sigma}{\sqrt{n}}$$

$$58.57 - 1.96\frac{1.175}{37.97} \le M(x) \le 58.57 - 1.96\frac{1.175}{37.97}$$

$$58.5 \le M(x) \le 58.64$$

t – табличне значення розподілу Стьюдента з числом ступенів свободи k і довірчою ймовірністю P