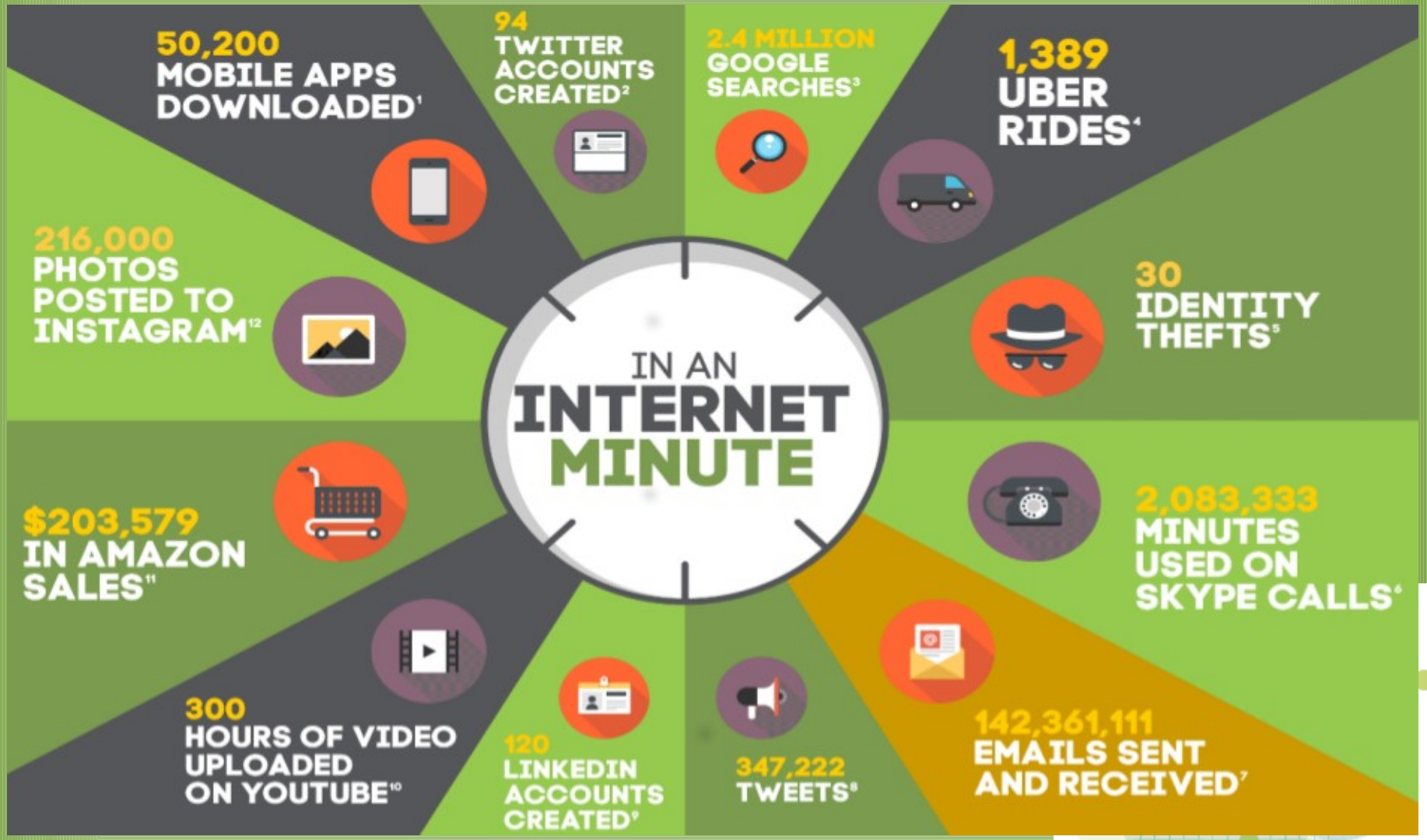
# Лекція 1

## Аналіз даних. Основні поняття. Описова статистика.

.

К.ф.-м.н. Щестюк Н.Ю.


50,200
MOBILE APPS
DOWNLOADED[1]
94
TWITTER
ACCOUNTS
CREATED[2]
2.4 MILLION
GOOGLE
SEARCHES[3]
1,389
UBER
RIDES[4]
216,000
PHOTOS
POSTED TO
INSTAGRAM[12]
30
IDENTITY
THEFTS[5]
IN AN
INTERNET
MINUTE
$203,579
IN AMAZON
SALES[11]
2,083,333
MINUTES
USED ON
SKYPE CALLS[6]
300
HOURS OF VIDEO
UPLOADED
ON YOUTUBE[10]
120
LINKEDIN
ACCOUNTS
CREATED[9]
347,222
TWEETS[8]
142,361,111
EMAILS SENT
AND RECEIVED[7]

# обмін даними

- **Uber** – найбільша в світі служба таксі, не володіє жодним авто
- **Facebook** - найпопулярніший власник медіа, не створює жодного контенту
- **Alibaba** - retailer, нічого не виробляє
- **Airbnb** – сервіс короткострокової оренди житла, не володіє жодним помешканням

- сервіс для пошуку пари для шлюбу
- сукупний дохід > 1 мільярда доларів
- близько 4% шлюбів у США в 2012
- більше 33 млн користувачів у понад 150 країнах

# Основні компоненти аналізу даних


"There are three kinds of lies:
lies, damned lies, and statistics."
B. Disraeli

# СТАТИСТИКА -

**наука про збір, організацію, аналіз та трактування даних.**

**(оцінити варіативність та зменшити невизначеність)**

# За допомогою статистики вирішують наступні проблеми:

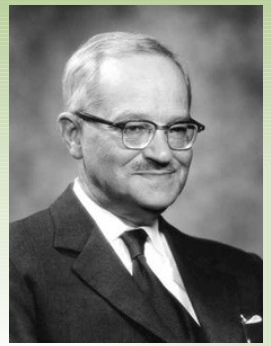
- оцінювання ризику вживання трансгенних продуктів харчування або нових вакцин
- передбачення кількості захворювань на грип за регіонами
- передбачення результатів наступних виборів
- управління самокерованими автомобілями
- голосові асистенти длч смартфонів

# Історія та нові напрямки розвитку статистики

Карл Гаусс
(1777-1855)

Карл Пірсон
(1857-1936),
Метод моментів,
критерій «хі-квадрат»

Рональд Фішер
(1890-1962)
Метод максимальної правдоподібності,
Fisher's exact test



Ежи Нейман
(1894-1977),
Біхевіористська статистика

А.М. Колмогоров
(1903-1987),
Основи непараметричної статистики

Статистика обєктів нечислової прир

Математичні методиплановання експерименту

Статистична теорія рішень

озробка пакетів прикладних програм аналізу даних

# Вибіркове обстеження (Dalenius,1974)

Потреба в статистичній інформації в предметній області

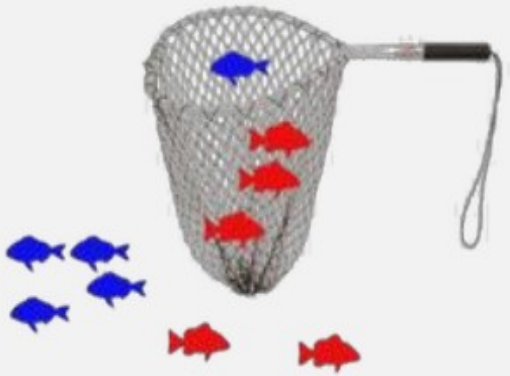
План вибіркового обстеження
(методи, обсяг вибірки)

Збір даних

Обробка, очищення та аналіз даних
(побудова моделей, перевірка гіпотез)

Трактування, презентація та використання результатів

# Генеральна сукупність



Генеральна сукупність – усі об'єкти, які хотів би вивчати дослідник при необмеженій кількості ресурсів.

# Як сформувати вибірку

**Простий випадковий вибір**
Всі об'єкти мають однакову можливість бути вибраними. Випадковим чином обирається n об'єктів

**Вибір з заміною**
Після того, як об'єкт вибрано, він повертається і може бути обраний повторно

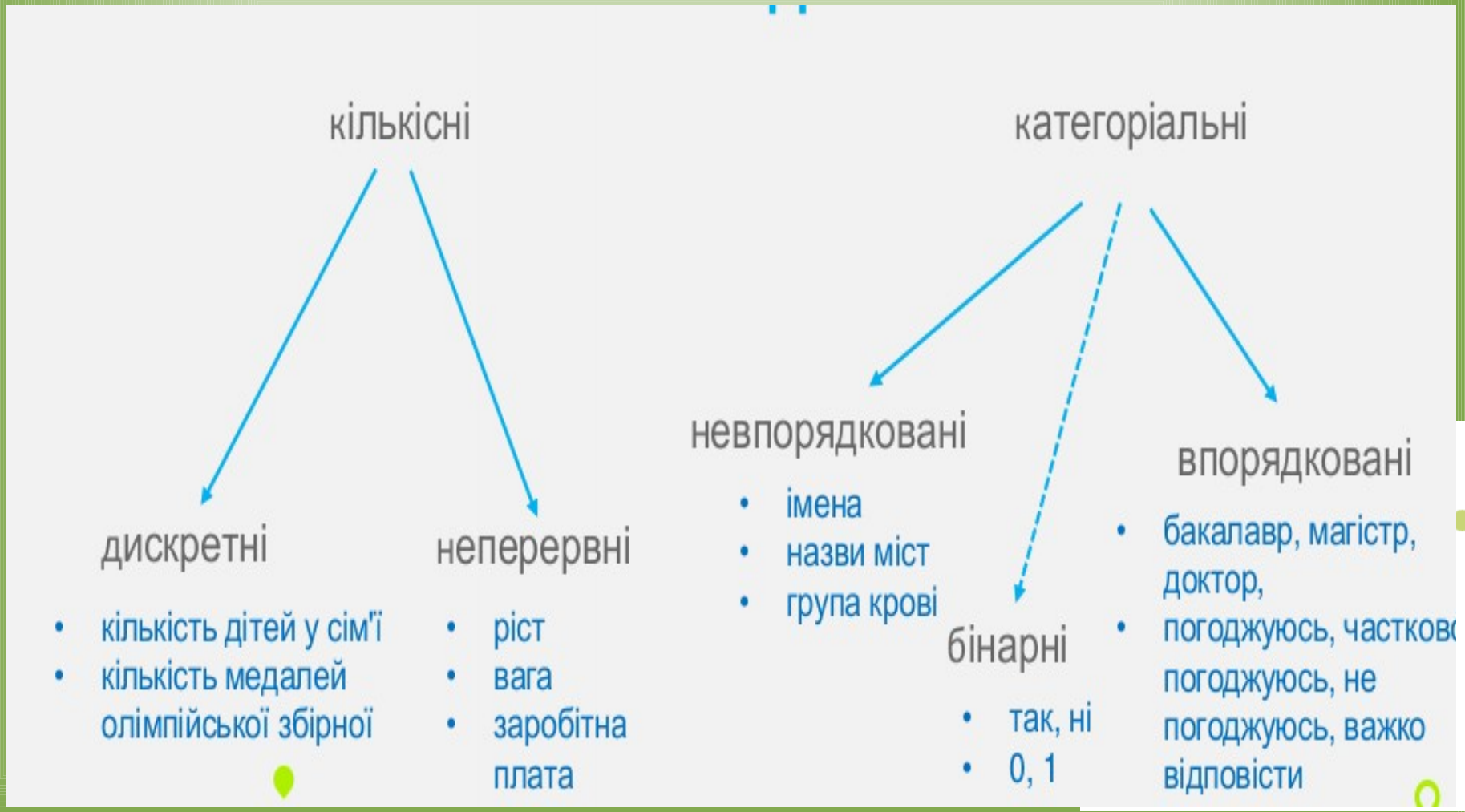
**Вибір без заміни**
Після того, як об'єкт вибрано, він вилучається і не може бути обраний повторно

**Стратометричний вибір**
Сукупність ділиться на гомогенні групи (населення за рівнем освіти чи віковою групою)

**Кластерний вибір**
Сукупність ділиться на кластери (місто на райони)

**Систематичний вибір**
Елементи сукупності впорядковуються і вибирається кожен k-ий елемент (елементи на конвейері з метою виявлення дефектів)

# Типи даних


кількісні
категоріальні
дискретні
неперервні
невпорядковані
впорядковані
бінарні
• кількість дітей у сім'ї
• кількість медалей олімпійської збірної
• ріст
• вага
• заробітна плата
• імена
• назви міст
• група крові
• так, ні
• 0, 1
• бакалавр, магістр, доктор,
• погоджуюсь, частково погоджуюсь, не погоджуюсь, важко відповісти

# Основні поняття

Генеральна сукупність

Вибірка
(n-об'єм)

Варіаційний ряд

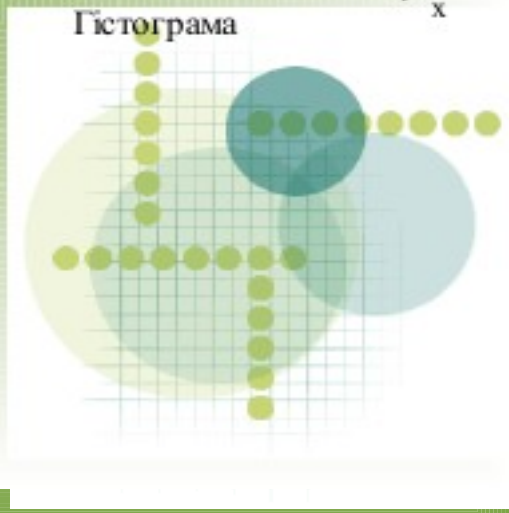

Полігон

Статистичний розподіл

| $x_i$ | $x_1$ | $x_2$ | $x_3$ | | $x_{k-1}$ | $x_k$ |
|---|---|---|---|---|---|---|
| $n_i$ | $n_1$ | $n_2$ | $n_3$ | | $n_{k-1}$ | $n_k$ |


Гістограма

Емпірична функція розподілу

# Вибірковий метод

**Генеральна сукупність** — це множина всіх значень, яких може набувати дана випадкова величина.

$\{x_1,\ldots,x_n\}$ – це **вибірка** спостережень за випадковою величиною $\xi$ із генеральної сукупності (*набір з незалежних і однаково розподілених випадкових величин ("копій")).*

*n* – **обсяг (об'єм)вибірки**, кількість елементів вибірки

Впорядкуємо вибірку за зростанням:

$$x_1^* \le x_2^* \le .. \le .x_n^*$$

Таким чином впорядковану вибірку

називають **варіаційним рядом**.

# Статистичний ряд

**Статистичним рядом** називається послідовність пар $(z_i, n_i)$, $i = \overline{1,l}$

**Груповані статистичний ряд (Frequency tables)** – це сукупність пар $(x_i^*, n_i^*)$, $i = \overline{1,k}$, де $x_i^*$ – середина і-того інтервалу,

$n_i^*$ – частота попадання у і-й інтервал.

Одна з рекомендацій щодо вибору числа інтервалів групування – формула Стерджесса:

$$k = 1 + \log_2 n$$

# ВІЗУАЛІЗАЦІЯ
# Полігон та гістограма частот

**Полігон частот** – ламана з вершинами у точках

$$(x_i^*, n_i^*),\ i = \overline{1,k}$$

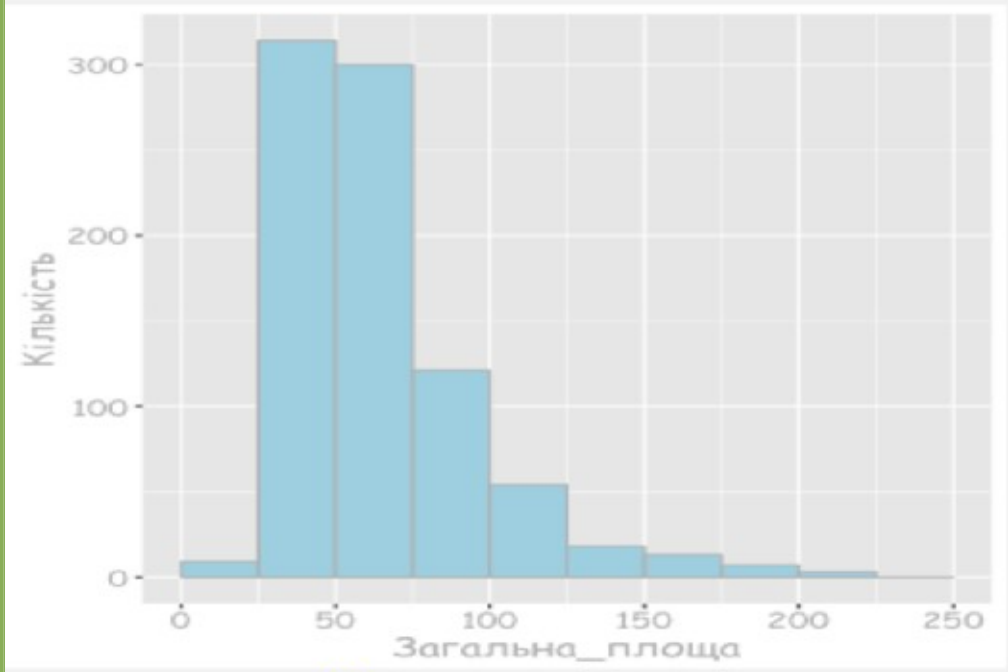
**Гістограма частот** – це ступінчаста фігура, побудована з прямокутників, основами яких є інтервали групування, а висоти

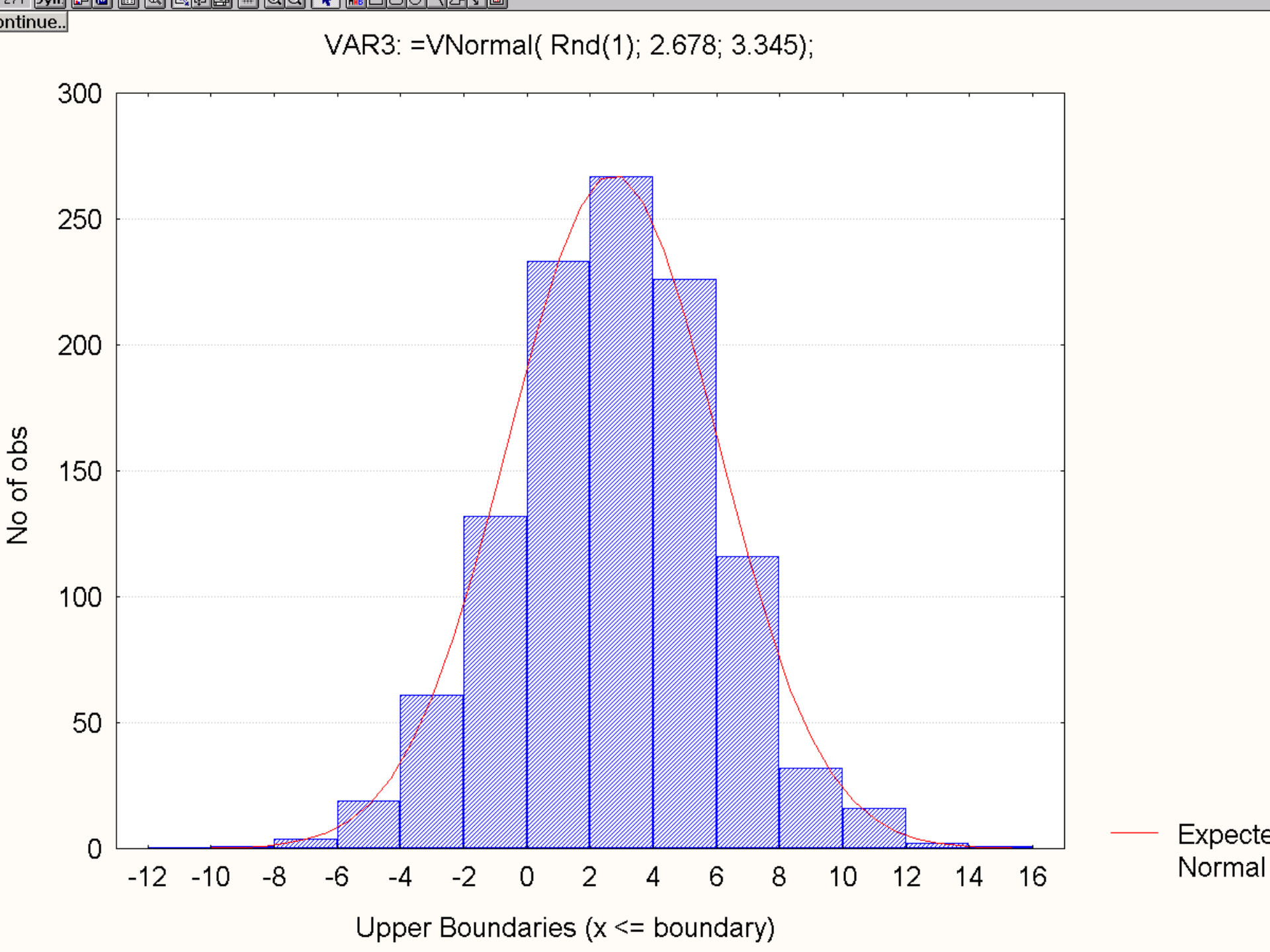
$$h_i^* = \frac{n_i^*}{\delta n},$$

де $\delta$ – це є довжина інтервалу групування.

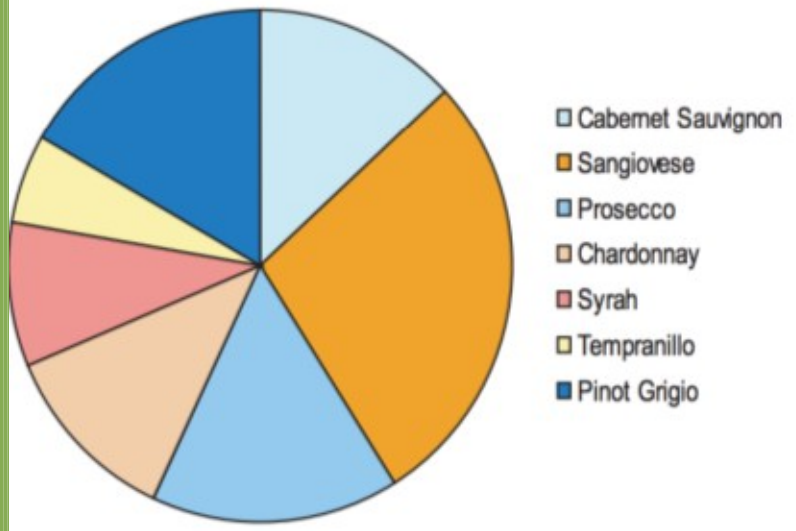
Площа гістограми дорівнює 1.

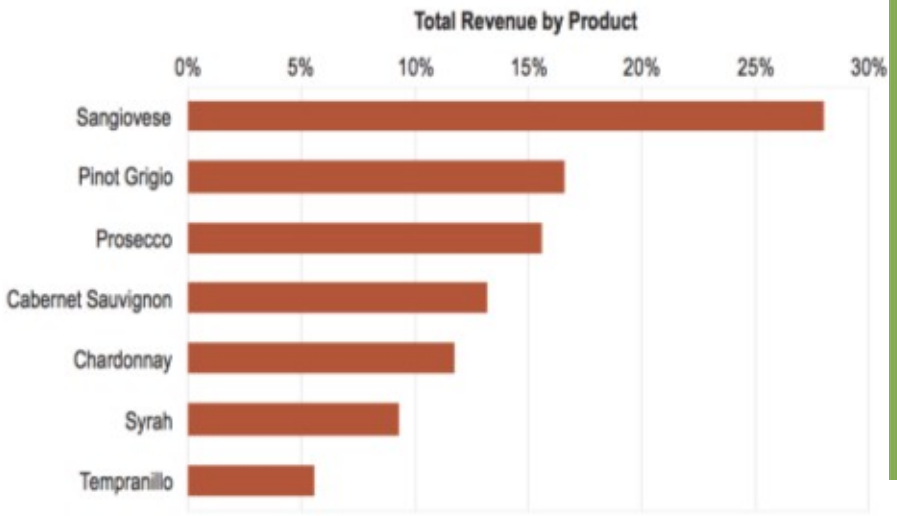
# гістограма



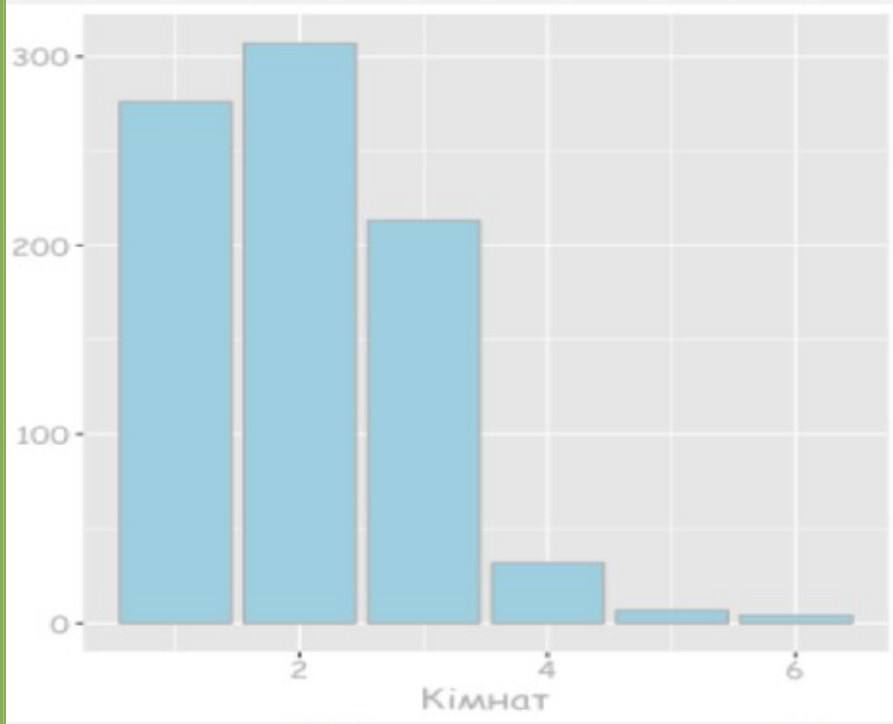

VAR3: =VNormal( Rnd(1); 2.678; 3.345);
No of obs
300
250
200
150
100
50
0
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
14
16
Upper Boundaries (x <= boundary)
Expecte
Normal

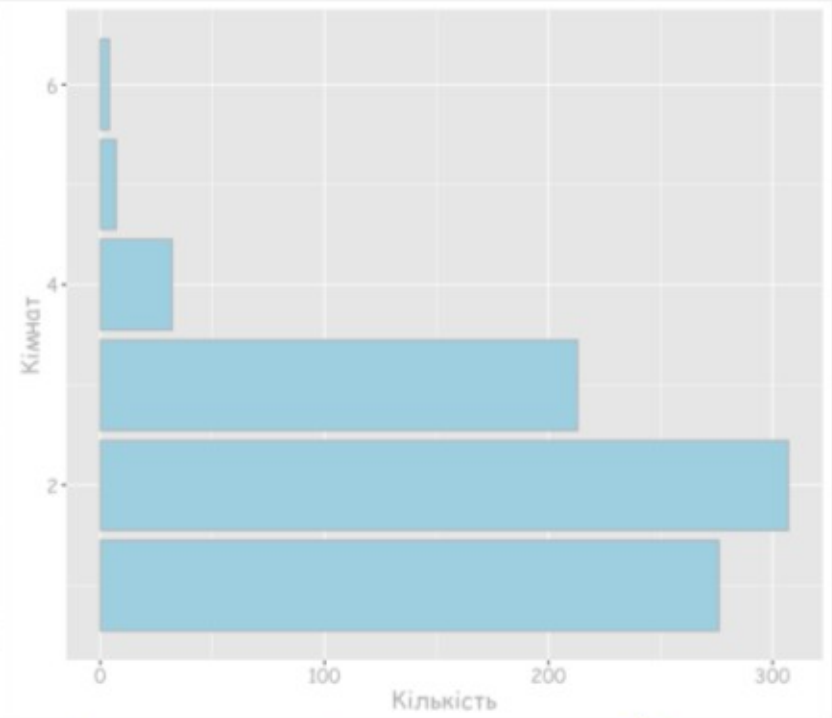
# кругова діаграма



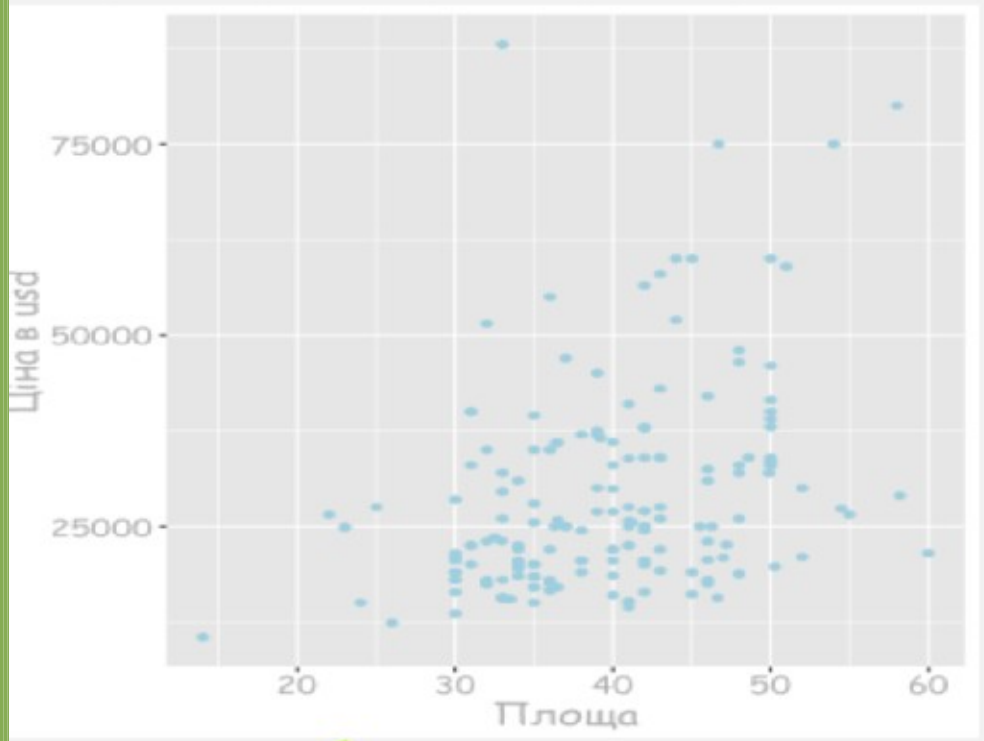



# стовпчикова діаграма

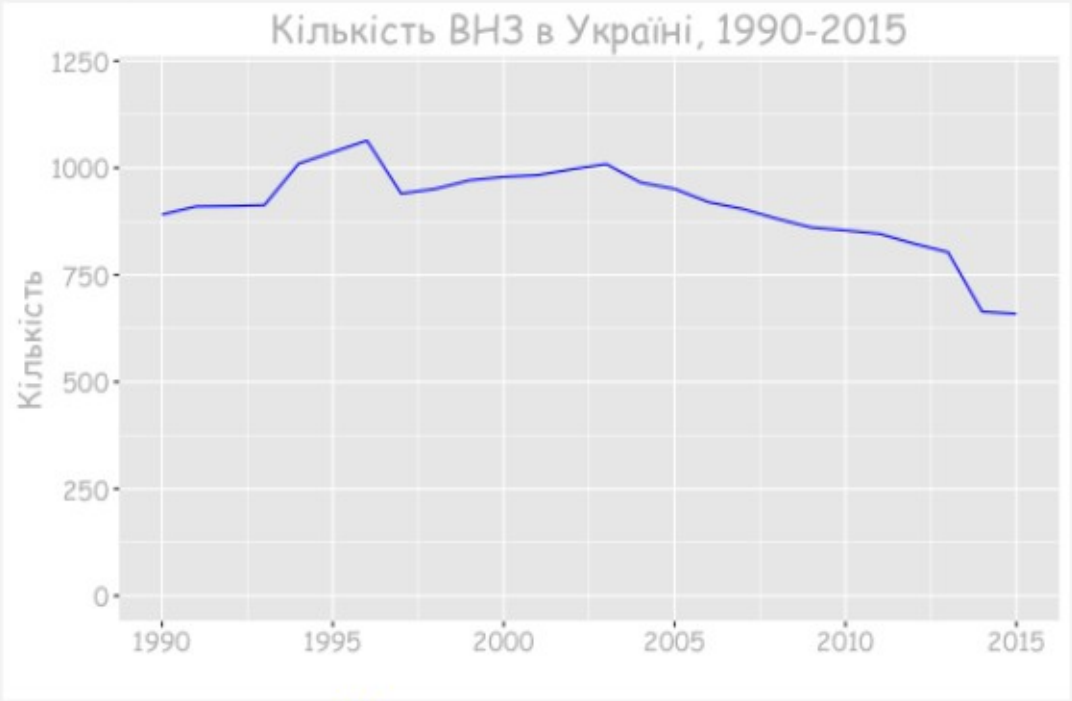


# діаграма розсіювання



# лінійний графік



# Емпірична функція

**Емпірична функція розподілу** визначається через статистичний ряд співвідношенням

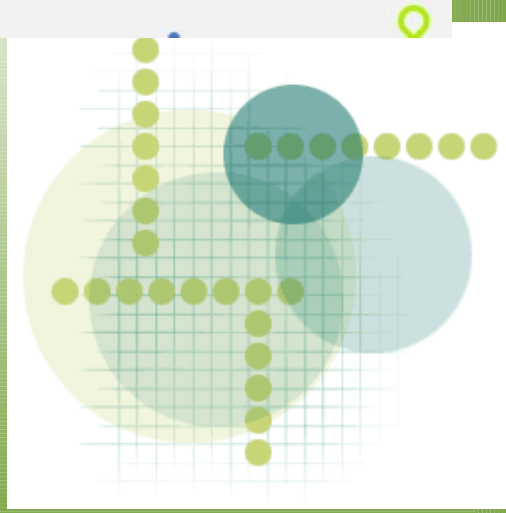
$$F_n^*(y) = \frac{1}{n}\sum_{i=1}^{n} I(x_i < y), \quad F_n^*(y) = \begin{cases} 0, & y \le x_{(1)}; \\ \frac{n_1}{n}, & x_{(1)} < y \le x_{(2)}; \\ \dots & \dots \\ \frac{n_1 + n_2 + \dots + n_k}{n}, & x_{(k)} < y \le x_{(k+1)}; \\ \dots & \dots \\ 1, & y > x_{(n)}. \end{cases}$$

**Кумулятивна крива** – ламана з вершинами у точках

$$\left( x_i^* + \frac{\delta}{2}, \frac{1}{n}\sum^{i} n_j^* \right), \; i = 1, \dots k$$

# який тип діаграми краще застосовувати?

- Порівнювати значення: стовпчикова діаграма, лінійний графік, графік розсіювання
- Зрозуміти композицію: (виділити складові) - стовпчикова діаграма, кругова діаграма
- Оцінити розподіл даних: лінійний графік, графік розсіювання, стовпчикова діаграма, гістограма
- Зрозуміти тренд: лінійний графік, стовпчикова діаграма
- Зрозуміти відношення між даними: лінійний графік, графік розсіювання



# Характеристики центральної тенденції

**Середнє значення**

$$\bar{x} = \frac{\sum x}{N}$$

Середня ціна однокімнатної квартири $25880, середня площа – 40 м²

**Медіана**
**ділить вибірку навпіл**

Медіана ціни однокімнатної квартири $23000, медіана площі 40 м²

**Мода**
**найчастіше трапляється**

# Характеристики центральної тенденції

**Вибіркове середнє:** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

**Вибіркове середнє для згрупованої вибірки:** $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^*$ ***де k – число інтервалів групування***

**Мода** (*Md*) – це елемент статистичного ряду з найбільшою частотою.

**Медіана (*Me*)-** це значення середнього елемента вибірки

**Середнє геометричне -** $\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

# середнє значення

3, 2, 5, 6, 7, 2, 3, 3

$$\bar{x} = \frac{3+2+5+6+7+2+3+3}{8} = \frac{31}{8} = 3.875$$

$$\bar{x} = \frac{\sum x}{N}$$

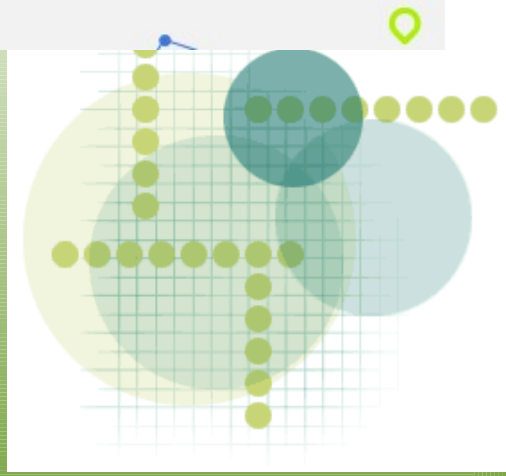
| Місце серед членів ООН | Місце серед усіх територій | Країна | Загальна очікувана тривалість життя при народженні | Чоловіча очікувана тривалість життя при народженні | Жіноча очікувана тривалість життя при народженні |
|---|---|---|---|---|---|
| | 1 | Макао | 84,36 | 81,39 | 87,47 |
| 1 | 2 | Андорра | 82,51 | 80,33 | 84,84 |
| 2 | 3 | Японія | 82,12 | 78,8 | 85,62 |
| 3 | 4 | Сінгапур | 81,98 | 79,37 | 84,78 |
| 4 | 5 | Сан-Марино | 81,97 | 78,53 | 85,72 |
| | 6 | Гонконг | 81,86 | 79,16 | 84,79 |
| 5 | 7 | Австралія | 81,63 | 79,25 | 84,14 |
| 6 | 8 | Канада | 81,23 | 78,69 | 83,91 |
| 7 | 9 | Франція | 80,98 | 77,79 | 84,33 |
| 8 | 10 | Швеція | 80,86 | 78,59 | 83,26 |

# медіана

| | |
|---|---|
| 3, 2, 5, 6, | 7, 2, 4, 3 |
| 2, 2, 3, 3, | 4, 5, 6, 7 |

$$\frac{3+4}{2} = 3.5$$

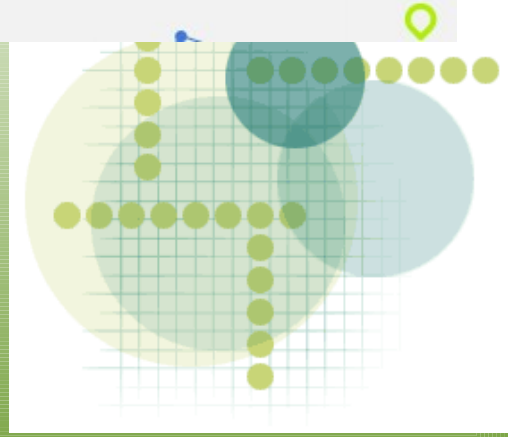
значення, яке ділить вибірку навпіл



# мода

3, 2, 5, 6, 7, 2, 3, 3
2, 2, 3, 3, 3, 5, 6, 7

значення, яке найчастіше трапляється



# Квантилі

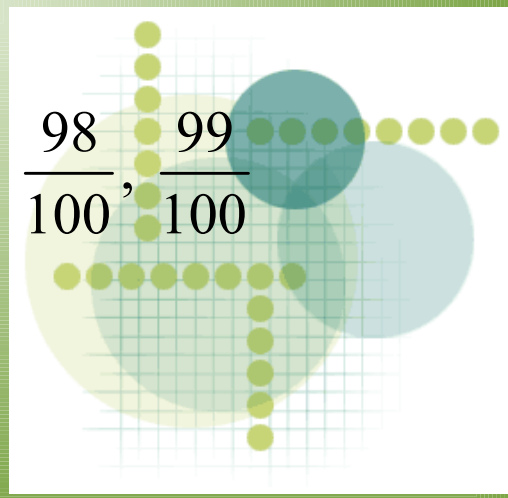
**Вибірковий квантиль** порядку $p$ – це абсциса точки, яка лежить на кумулятивній кривій та має ординату $p$.

**Медіана** – це квантиль порядку $\frac{1}{2}$ .

**Квартилі** – квантилі порядку $\frac{1}{4}, \frac{2}{4}$ .та $\frac{3}{4}$

**Децилі** – квантилі порядку $\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}$

**Процентилі** – квантилі порядку $\frac{1}{100}, \frac{2}{100}, \ldots, \frac{98}{100}, \frac{99}{100}$



# Характеристики розсіювання

**Розмах (range) вибірки (ϖ)**, тобто різниця між найбільшим та найменшим елементами вибірки.

Дозволяє оцінити розкид елементів у вибірці.

**Міжквартильний розмах** (inter-quartile range або mid-spread) – це є різниця між нижнім і верхнім квартилем.

Дозволяє оцінити розкид 50% елементів вибірки і не враховує вплив викидів.

# Характеристики розсіювання

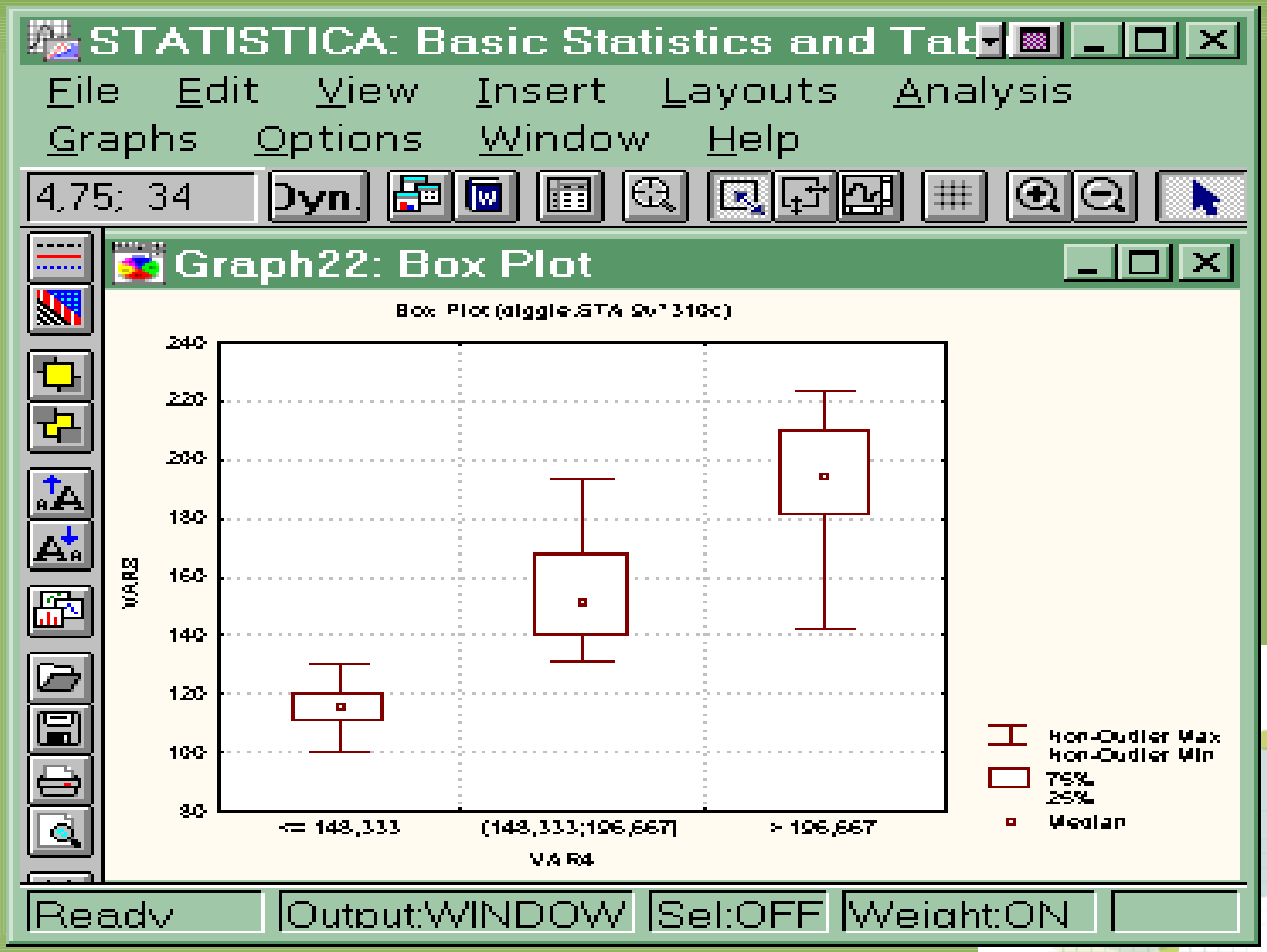
**Вибіркова дисперсія —** $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

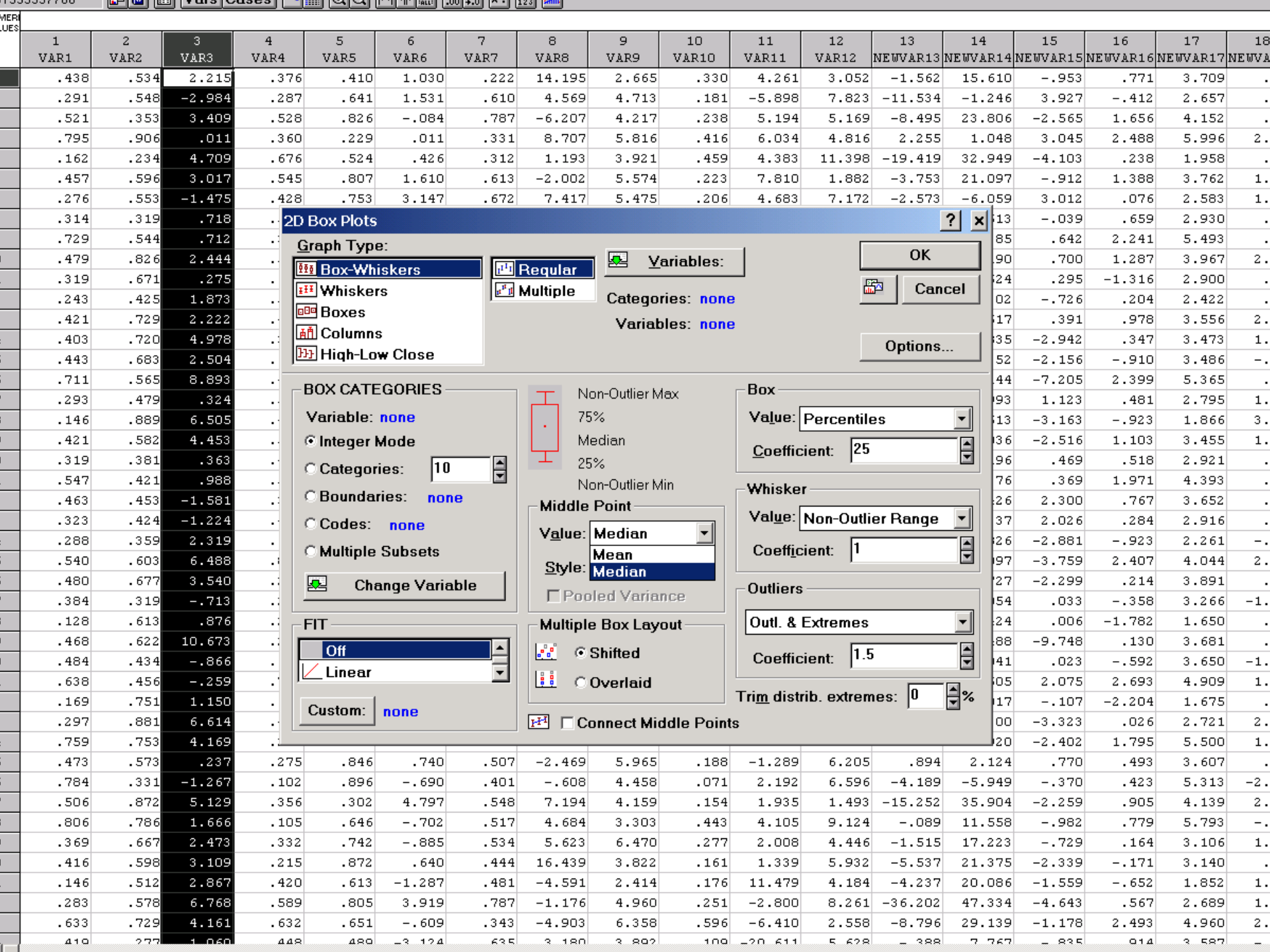
**Для групованих даних —** $s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i^* - \bar{x})^2$

$$s_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

**Незміщена вибіркова дисперсія**

**Середньоквадратичне відхилення – це квадратний корінь з вибіркової дисперсії:** $\sigma_B = \sqrt{s^2}$


STATISTICA: Basic Statistics and Tab
File Edit View Insert Layouts Analysis
Graphs Options Window Help
4,75; 34
Dyn.
Graph22: Box Plot
Box Plot (diggle.STA 9v*310c)
240
220
200
180
160
140
120
100
80
VAR3
<= 148,333
(148,333;196,667]
> 196,667
VAR4
Non-Outlier Max
Non-Outlier Min
75%
25%
Median
Ready
Output:WINDOW
Sel:OFF
Weight:ON


2D Box Plots
Graph Type:
Box-Whiskers
Whiskers
Boxes
Columns
High-Low Close
Regular
Multiple
Variables:
Categories: none
Variables: none
OK
Cancel
Options...
BOX CATEGORIES
Variable: none
Integer Mode
Categories: 10
Boundaries: none
Codes: none
Multiple Subsets
Change Variable
Non-Outlier Max
75%
Median
25%
Non-Outlier Min
Middle Point
Value: Median
Mean
Style: Median
Pooled Variance
FIT
Off
Linear
Custom: none
Multiple Box Layout
Shifted
Overlaid
Connect Middle Points
Box
Value: Percentiles
Coefficient: 25
Whisker
Value: Non-Outlier Range
Coefficient: 1
Outliers
Outl. & Extremes
Coefficient: 1.5
Trim distrib. extremes: 0 %

# Як вибирати показник розсіювання?

1. Якщо за характеристику центральної тенденції вибрано медіану, то слід вибрати міжквартильний розмах.

2. Якщо за характеристику центральної тенденції вибрано вибіркове середнє, то слід вибрати дисперсію та середньоквадратичне відхилення.

# Коефіцієнт асиметрії

**Коефіцієнт асиметрії** (skewness) обчислюється як:

$$a_s = \frac{m_3}{\sigma_B^3} \qquad m_3 = \frac{1}{n}\sum n_i (x_i - \bar{x})^3$$

Якщо коефіцієнт асиметрії **додатній**:

- Розподіл має довгий правий хвіст
- Вибіркове середнє більше за медіану
- Медіана більша за моду

Якщо коефіцієнт асиметрії **від'ємний**:

- Розподіл має довгий лівий хвіст
- Вибіркове середнє менше за медіану
- Медіана менша за моду

# Коефіцієнт ексцесу

*Вибірковим коефіцієнтом ексцесу (kurtosis) називається величина*

$$e_k = \frac{m_4}{\sigma_B^4} - 3 \qquad m_4 = \frac{1}{n}\sum n_i (x_i - \bar{x})^4$$

Якщо коефіцієнт ексцесу **< 3** , то пік розподілу пологіший, ніж у нормального – розподіл **плосковершинний**.

Якщо коефіцієнт ексцесу **> 3**, то пік розподілу крутіший, ніж у нормального – розподіл **гостровершинний**.

# ІНТЕРПРЕТАЦІЯ РЕЗУЛЬТАТІВ

## парадокс Сімпсона

| | | Подало заяв | Прийнято | Відсоток прийнятих |
|---|---|---|---|---|
| Факультет А | Чоловіки | 900 | 450 | 50% |
| | Жінки | 100 | 80 | 80% |
| Факультет Б | Чоловіки | 100 | 10 | 10% |
| | Жінки | 900 | 180 | 20% |

# парадокс Сімпсона

| | | Подало заяв | Прийнято | Відсоток прийнятих |
|---|---|---|---|---|
| Факультет А | Чоловіки | 900 | 450 | 50% |
| | Жінки | 100 | 80 | 80% |
| Факультет Б | Чоловіки | 100 | 10 | 10% |
| | Жінки | 900 | 180 | 20% |
| Обидва | Чоловіки | 1000 | 460 | 46% |
| | Жінки | 1000 | 260 | 26% |

# Методика та філософія викладання

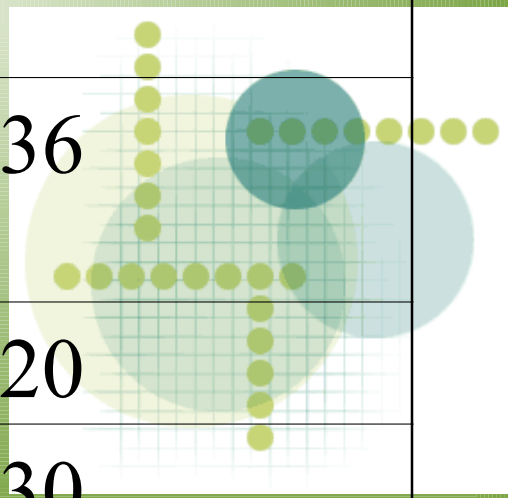
Теоретичний матеріал подається у вигляді презентацій та викладень на дошці

Засвоєння теоретичного матеріалу відбувається за рахунок участі у дискусіях, тестуванні, колоквіуму

Практичні навички формуються при виконанні індивідуальних завдань

# Умови визначення навчального рейтингу

| № | Вид роботи | Кількість | Макс. кількість балів |
|---|---|---|---|
| 1 | Дискусія (тест) на лекціях | 7 | 14 |
| 2 | Участь у проекті | 3 | 36 |
| 3 | колоквіум | 1 | 20 |
| 4 | Ісит | 1 | 30 |



# Дякую за увагу!