Київський національний університет

імені Тараса Шевченка

Звіт

до лабораторної роботи 6,

дисципліни: “Технології аналізу та візуалізації даних”,

студента: Євчика Олексія,

групи Інформатика, Магістри

Тема: Візуалізація розподілених даних з використанням гістограм, ядерної оцінки щільності (KDE) та діаграм розмахів (Boxplot) (R).

Мета роботи:

1. Засвоїти принципи відображення розподілу даних за допомогою гістограм, KDE та boxplot.
2. Навчитися налаштовувати параметри бінінгу, ядра та «вусів» для коректної інтерпретації.
3. Оволодіти базовими прийомами попередньої обробки, виявлення викидів і побудови порівняльних візуалізацій у R.

Ознайомитися з методами візуалізації розподілених даних у R, навчитися застосовувати гістограми, ядерну оцінку щільності та діаграми розмахів, дослідити вплив параметрів (binwidth, bandwidth, adjust, правила визначення вусів) на форму графіків та інтерпретацію даних.

Теоретична частина

1. Гістограма

Ідея. Гістограма є методом візуалізації розподілу числової змінної шляхом дискретизації осі значень на інтервали (біни) та підрахунку кількості спостережень або оцінки щільності у кожному біні.

Ключові параметри.

- *binwidth* — ширина біну;
- *breaks / number of bins* — кількість та розташування інтервалів;
- *limits* — межі відображення;
- *count vs density* — шкала частот або нормована щільність.

Зауваги. Занадто широкий бін може приховати багатомодальність, тоді як надто вузький — створює шум та “зубчастість” графіка.

2. Ядерна оцінка щільності (KDE)

Ідея. KDE — непараметричний метод оцінки щільності ймовірності, що базується на згладженні спостережень за допомогою ядра (частіше Gaussian).

Параметри згладження.

- *bandwidth* (ширина вікна) визначає ступінь згладження;
- у ggplot2 масштабування bandwidth задається через аргумент *adjust*.

Компроміс bandwidth.

- більший bw → гладка крива, але втрата локальних деталей;
- менший bw → нерівний контур і чутливість до шуму.

3. Діаграма розмахів (Boxplot)

Складові.

- медіана (лінія усередині коробки),
- межі квартилів Q1–Q3 (коробка),
- “вуса” (зазвичай до 1.5 IQR або іншого правила),
- точки-викиди (спостереження за межами вусів).

Призначення. Порівняння розподілів між групами, робастне відображення центральної тенденції та розкиду.

Варіації.

- boxplot із накладеними точками (stripplot, jitter);

- violin plot для візуалізації форми щільності.

4. Порівняння методів

- **Гістограма:** дає загальну форму розподілу, але залежить від вибору бінів.
- **KDE:** гладка безбінова оцінка, чутлива до bandwidth.
- **Boxplot:** відображає позиційні характеристики та викиди, але не показує детальну форму розподілу.

---

5. Filled Contour (залиті контури)

5.1. Ідея методу

Filled contour — це спосіб візуалізації *двовимірної поверхні* (z = f(x, y)) у вигляді рівневих ліній із заливкою між рівнями. Метод дозволяє відобразити структуру просторових даних, теплових карт, топографії та поверхонь щільності.

Графік формує набір рівнів (levels), між якими простір зафарбовується окремими кольорами, що полегшує сприйняття градієнтів і переходів значень.

5.2. Вибір палітри

Колірна палітра є критичною для коректної інтерпретації таких карт.
Основні вимоги до палітри для filled contour:

- **послідовність (perceptual uniformity)** — однаковий крок кольору відповідає однаковій зміні значення;

- **дружність до дальтонізму** — кольори повинні бути розрізнюваними навіть у випадках порушення колірного зору;
- **монотонність яскравості** — щоб забезпечити правильне сприйняття “високих” і “низьких” значень;
- **читабельність у друці та в градаціях сірого.**

Палітри сімейства **viridis** (viridisLite::viridis) повністю задовольняють ці вимоги. Вони були спеціально розроблені для наукової візуалізації та є стандартом де-факто у багатьох бібліотеках.

3. Дані та попередня обробка

У даній роботі були використані випадково згенеровані дані.

Попередня обробка включала:

- **фільтрацію** (видалення некоректних або явно помилкових записів),
- **масштабування** (за потреби — стандартне масштабування або нормалізація),
- **обробку викидів**: попередній аналіз за допомогою boxplot; рішення — видалення / заміна / залишення без змін, залежно від мети аналізу.
- **приведення типів даних** (числові, фактори, дати).

4. Виявлення викидів за IQR-правилом

4.1. Мета

Мета індивідуального завдання — ідентифікувати викиди у вибраній числовій змінній з використанням IQR-правила (Interquartile Range Rule), визначити порогові значення та проаналізувати природу знайдених викидів.

4.2. Теоретичні відомості

IQR-правило ґрунтується на інтерквартильному розмаху, що визначається як:

$$IQR = Q3 - Q1,$$

де:

- **Q1** — перший квартиль (25-й перцентиль),
- **Q3** — третій квартиль (75-й перцентиль).

Порогові значення для виявлення викидів визначаються як:

$$\text{Lower Bound} = Q1 - 1.5 \cdot IQR,$$

$$\text{Upper Bound} = Q3 + 1.5 \cdot IQR.$$

Будь-які спостереження, що лежать нижче нижнього порогу або вище верхнього, вважаються потенційними викидами.

Переваги методу: простота, робастність до шуму, нечутливість до великих амплітудних викидів.
Недоліки: IQR-правило добре працює для одномодальних симетричних розподілів, але може давати багато "штучних" викидів у сильно асиметричних розподілах.

4.3. Кроки виконання

Обчислення квартилів:
Отримано значення Q1 та Q3 для обраної змінної.

Обчислення IQR:

$$IQR = Q3 - Q1$$

1. **Визначення порогів:**

   - нижній поріг = Q1 − 1.5·IQR
   - верхній поріг = Q3 + 1.5·IQR

2. **Підрахунок кількості викидів:**

   Визначено кількість спостережень нижче та вище порогів.

3. **Формування таблиці результатів та коротка інтерпретація.**

6.4. Очікуваний результат

Нижче наведено загальний формат таблиці (числа заповнюються після обчислень у R).

Код:

```
## ------------------------------------------------------------
## 1. Підготовка середовища
## ------------------------------------------------------------
packages <- c("ggplot2","dplyr","tidyr","readr","scales","ggpubr")
to_install <- setdiff(packages, rownames(installed.packages()))
if(length(to_install)) install.packages(to_install)
lapply(packages, library, character.only = TRUE)

set.seed(123)
```

```r
## ------------------------------------------------------------
## 2. Дані-примірники
## ------------------------------------------------------------

# 1) Вбудований набір даних
data("faithful")   # колонки: eruptions, waiting

# 2) Синтетика
x1 <- rnorm(400, mean = 0, sd = 1)
x2 <- rnorm(300, mean = 3, sd = 0.7)
df_mix <- data.frame(x = c(x1, x2))
```

```r
## 3. Гістограми: ширина біну, count vs density
## ------------------------------------------------------------

# Count
p_h1 <- ggplot(faithful, aes(x = eruptions)) +
  geom_histogram(binwidth = 0.2, fill = "grey70", color = "grey30") +
  labs(title = "Гістограма (лічильник)",
       x = "Тривалість виверження (хв)", y = "Кількість") +
  theme_minimal(base_size = 12)

# Density
p_h2 <- ggplot(faithful, aes(x = eruptions, y = after_stat(density))) +
  geom_histogram(binwidth = 0.2, fill = "grey70", color = "grey30") +
  labs(title = "Гістограма (щільність)",
       x = "Тривалість (хв)", y = "Щільність") +
  theme_minimal(12)

# Вплив binwidth — малий
p_bw_small <- ggplot(df_mix, aes(x)) +
  geom_histogram(binwidth = 0.15, fill = "steelblue", color = "white") +
  labs(title = "Малий binwidth (деталі + шум)", x = "x", y = "Count") +
  theme_minimal(12)
```

```r
# Вплив binwidth — великий
p_bw_large <- ggplot(df_mix, aes(x)) +
  geom_histogram(binwidth = 0.6, fill = "steelblue", color = "white") +
  labs(title = "Великий binwidth (згладження + ризик втрати мод)", x = "x", y = "Count") +
  theme_minimal(12)
```

```r
## 4. KDE та накладання на гістограму
## ------------------------------------------------------------

p_kde <- ggplot(df_mix, aes(x)) +
  geom_density(linewidth = 1) +
  labs(title = "KDE (оцінка щільності)", x = "x", y = "Щільність") +
  theme_minimal(12)

p_hist_kde <- ggplot(df_mix, aes(x, y = after_stat(density))) +
  geom_histogram(binwidth = 0.3, fill = "grey80", color = "grey40") +
  geom_density(linewidth = 1) +
  labs(title = "Гістограма + KDE", x = "x", y = "Щільність") +
  theme_minimal(12)

# Вплив adjust
p_kde_adj06 <- ggplot(df_mix, aes(x)) +
  geom_density(adjust = 0.6, linewidth = 1) +
  labs(title = "KDE: adjust = 0.6 (детальніше)", x = "x", y = "Щільність") +
  theme_minimal(12)

p_kde_adj10 <- ggplot(df_mix, aes(x)) +
  geom_density(adjust = 1.0, linewidth = 1) +
  labs(title = "KDE: adjust = 1.0 (баланс)", x = "x", y = "Щільність") +
  theme_minimal(12)

```

```r
## ------------------------------------------------------------------
## 5. Boxplot + IQR-викиди
## ------------------------------------------------------------------

p_box <- ggplot(iris, aes(x = Species, y = Sepal.Length)) +
  geom_boxplot(outlier.colour = "red", width = 0.6) +
  geom_jitter(width = 0.1, alpha = 0.4) +
  labs(title = "Boxplot: Sepal.Length за видами",
       x = "Вид", y = "Sepal.Length") +
  theme_minimal(12)
```

```r

## --- Визначення викидів за IQR ---
x <- faithful$eruptions

Q1  <- quantile(x, 0.25)
Q3  <- quantile(x, 0.75)
IQRv <- IQR(x)

lower <- Q1 - 1.5 * IQRv
upper <- Q3 + 1.5 * IQRv

out_idx <- which(x < lower | x > upper)
out_values <- x[out_idx]

## Таблиця порогів IQR
iqr_table <- data.frame(
  Metric = c("Q1", "Q3", "IQR", "Нижній поріг", "Верхній поріг", "Кількість викидів"),
  Value  = c(Q1, Q3, IQRv, lower, upper, length(out_idx))
)
```

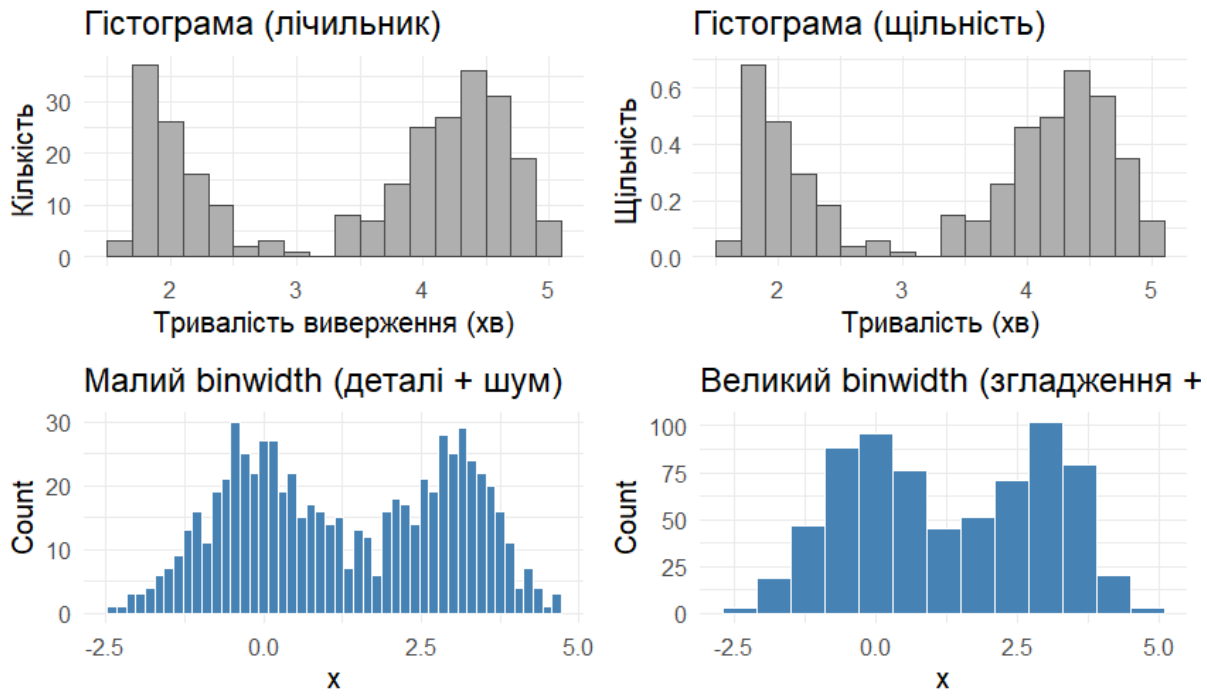
```r
## Таблиця порогів IQR
iqr_table <- data.frame(
  Metric = c("Q1", "Q3", "IQR", "Нижній поріг", "Верхній поріг", "Кількість викидів"),
  Value  = c(Q1, Q3, IQRv, lower, upper, length(out_idx))
)
```

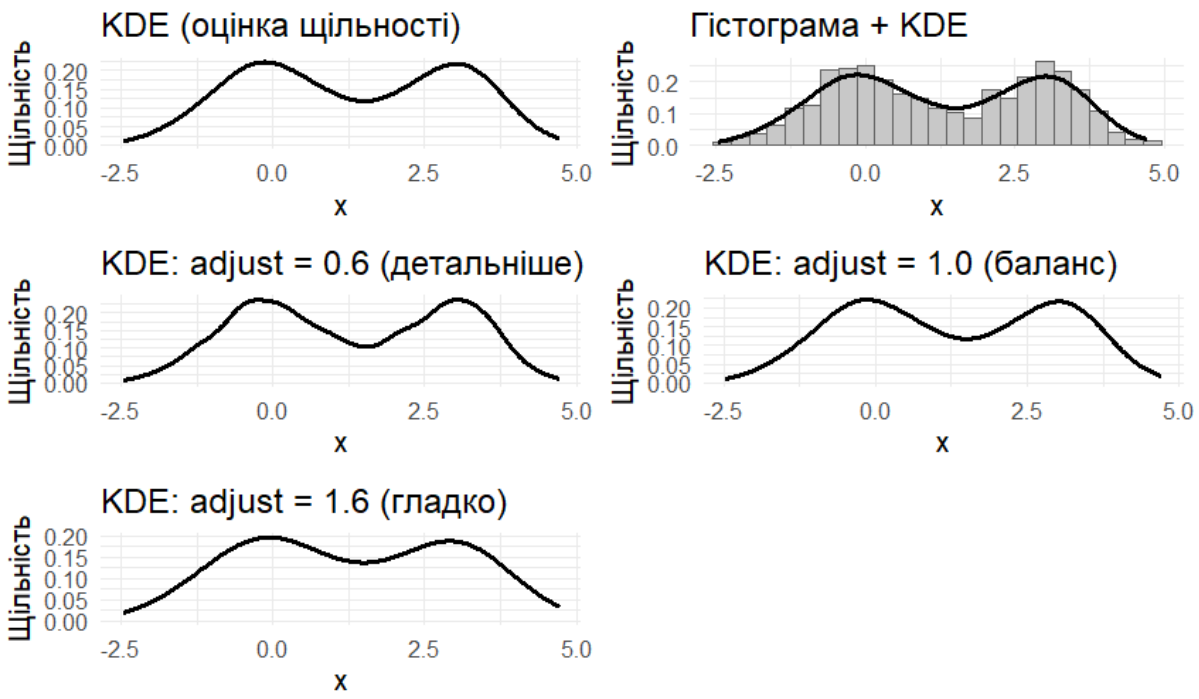
```r
## ------------------------------------------------------------
## 6. Компоновка фігур
## ------------------------------------------------------------

ggpubr::ggarrange(p_h1, p_h2, p_bw_small, p_bw_large, ncol = 2, nrow = 2)
ggpubr::ggarrange(p_kde, p_hist_kde, p_kde_adj06, p_kde_adj10, p_kde_adj16,
                  ncol = 2, nrow = 3)

## ------------------------------------------------------------
## 7. Збереження результатів
## ------------------------------------------------------------
ggsave("hist_density_overlay.png", p_hist_kde, width = 8, height = 5, dpi = 300)
ggsave("box_iris.png", p_box, width = 7, height = 5, dpi = 300)
```

Результат:





Контрольні запитання та відповіді:

1. Як вибір binwidth впливає на форму та інтерпретацію гістограми?

- Малий binwidth → гістограма стає «зашумленою», показує дрібні коливання, може створити хибні моди.
- Великий binwidth → занадто згладжує розподіл, приховує важливі структури (наприклад, мультимодальність).
- Тому binwidth визначає баланс між деталізацією та згладженістю.

2. У чому різниця між відображенням частоти (count) і щільності (density) на гістограмі?

- Count — показує абсолютну кількість спостережень у кожному біні.
- Density — масштабує площу гістограми до 1 (щоб можна було порівнювати з KDE та між вибірками різного розміру).

Density обов’язкова, якщо гістограму хочуть накладати на KDE.

3. Яку роль відіграє параметр adjust у KDE і як його обирати?

- adjust множить базову ширину ядра (bandwidth).
- Менше 1 → менше згладження, більше деталей, але більше шуму.
- Більше 1 → сильніше згладження, менше шуму, але можливе зникнення мод.
- Обирають за метою аналізу:
    - Для виявлення структур → 0.6–1.0
    - Для гладкої презентації → 1.2–1.6

4. Чому KDE чутлива до ширини ядра? Що таке компроміс «зміщення–дисперсія»?

- Малий bandwidth → низьке зміщення, висока дисперсія (занадто змінна крива).
- Великий bandwidth → високе зміщення, низька дисперсія (занадто згладжена крива).
- Компроміс зміщення–дисперсія означає, що не можна одночасно мати і низьке зміщення, і низьку дисперсію.

5. За яким правилом формуються «вуса» у boxplot за замовчуванням?

У класичному Tukey boxplot:

- Нижній вус = $Q1 - 1.5 \times IQR$
- Верхній вус = $Q3 + 1.5 \times IQR$
- Точки поза вусами вважаються потенційними викидами.

6. Чому boxplot вважається робастним до викидів?

Бо він використовує медіану та квартилі, а не середнє і стандартне відхилення.
Ці статистики стійкі: значення викидів не відсувають їх суттєво.

7. Коли доцільно комбінувати histogram + KDE на одному графіку?

- Коли потрібно одночасно показати:
    - емпіричний розподіл (гістограма)
    - гладке наближення розподілу (KDE)
- Особливо корисно при аналізі:
    - мультимодальності
    - порівняння вибірок
    - демонстрації різних bandwidth

8. Які підходи до виявлення та обробки викидів ви знаєте (IQR, z-score тощо)?

Методи виявлення:

- IQR-правило (Tukey)
- z-score: $|z| > 3$
- robust z-score (MAD)
- діаграми: boxplot, scatter + LOF
- кластеризація (DBSCAN)
- моделі зважування (robust regression)

Методи обробки:

- вилучення
- winsorizing (обрізання)

- лог-трансформації
- заміна на медіану/квартиль
- окреме моделювання груп

9. Які ризики неправильного масштабування осей/форматування підписів?

- Неправильне масштабування може створити візуальні викривлення, зокрема:
    - перебільшення різниць
    - приховання варіації
    - некоректне інтерпретування трендів
- Погані підписи призводять до:
    - неправильного розуміння одиниць
    - плутанини між density та count
    - хибних висновків у порівняннях

10. Як забезпечити відтворюваність графіків у звіті?

- фіксувати set.seed()
- зберігати версії пакетів (sessionInfo())
- зберігати всі дані/генерацію у файлі
- зберігати всі параметри графіків (binwidth, adjust, limits)
- використовувати RMarkdown або Quarto, де виконання коду повністю контрольоване
- уникати випадкових аспектів (джиттер з width, alpha без seed)