Course: Computer statistics

Lecture 2: Primary graphical data analysis.

Lecturer: Oleksandr Dykhovychnyi

Лекція 2. Первинний графічний аналіз даних.

I. **Емпірична функція розподілу**

II. **Гістограма**

III. **«Скриня з вусами» (box-whisker plots)**

IV. **P-P і Q-Q діаграми**

Розглянемо, як графічно зображують основні характеристики розподілу вибірки, і як саме це реалізується засобами мови **R.**

I. Емпірична функція розподілу

Для вибірки $X = (x_1, x_2, ..., x_n)$ **емпіричною функцію розподілу** називають наступну функцію:

$$F_n^*(x) = \frac{1}{n}\sum_{x_i<x} I_{(-\infty,x)}(x_k), \quad x \in (-\infty, \infty).$$

В **R** для вибірки X її будує функція **ecdf(x),** де **x -** вектор, матриця або фрейм.

Приклади: Згенеруємо 100 значень біноміально розподіленої випадкової величини з параметрами 5 і 0.5 і знайдемо для такої вибірки емпіричну функцію розподілу та побудуємо її графік.

```
> set.seed(0) #фіксуємо початкове значення генератора випадкових чисел
> x<-rbinom(100,5,0.5)  #генеруємо випадкові числа
   # будуємо графік
> plot(ecdf(x),col="blue",ylab="Емпірична функція розподілу")
```

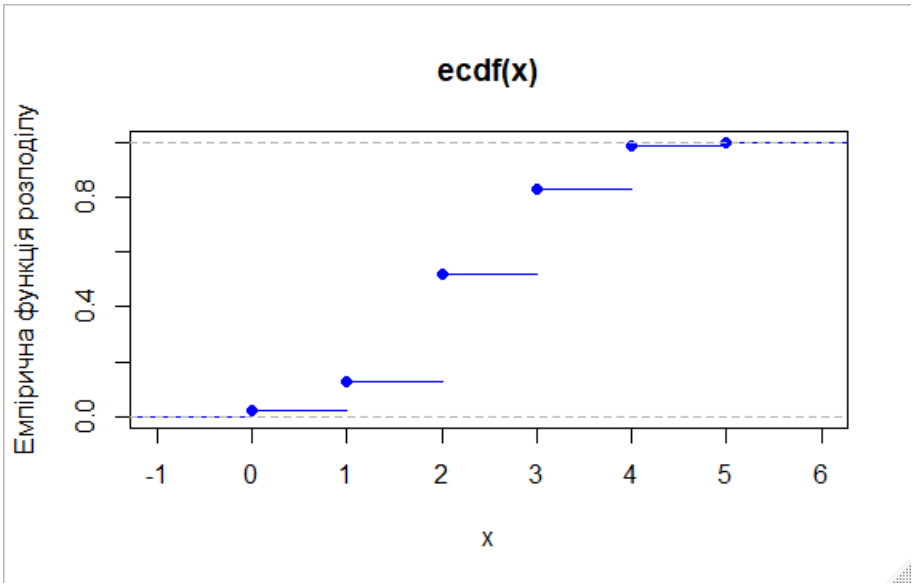

Рис.1(Малюнок згенеровано **RStudio**)

Функція **knots()** дозволяє визначити елементи вибірки, в яких відбуваються стрибки емпіричної функції розподілу.

Приклади:

```
> knots(ecdf(x))
[1] 0 1 2 3 4 5
```

II. Гістограма

В системі **R** для побудови **гістограми** використовують функцію **hist(x)**.

Параметри функції **hist(x**….):

- **x** —вектор, матриця або фрейм (вибірка).

- **breaks** — параметр, який відповідає за розбиття на інтервали.

Це може бути числовий вектор, що визначає межі розбиття; число, яке задає кількість інтервалів розбиття; символьна змінна, що визначає алгоритм розбиття на інтервали; функція, що обчислює кількість інтервалів розбиття.

-**freq** і **probability** — два логічних і альтернативних один одному аргументи (одночасно їх не рекомендовано задавати).

Якщо **freq = TRUE** (або **probability** = **FALSE**), то будується гістограма частот.
Якщо **freq = FALSE** (або **probability = TRUE**) — гістограма відносних частот.
По замовчуванню будується гістограма частот.

- **include.lowest** — логічний аргумент. Мінімальний елемент вибірки входить в перший інтервал в якості лівої границі. Використовується лише в тому випадку, коли **breaks** — числовий вектор.

- **right** — логічний аргумент. Якщо **right = TRUE**, то інтервали розбиття мають вигляд (a ; b].

- **density** і **angle** — числові аргументи, що відповідають за щільність і кут нахилу штриховки стовпчиків гістограми.

- **col** — символьний або числовий аргумент, що задає або колір стовпчиків гістограми (якщо не задані аргументи density і angle), або колір штриховки.

- **border** — колір границь стовпчиків гістограми.

- **main** — заголовок гістограми.

- **xlim і ylim** — границі осей гістограми.

- **xlab** і **ylab** — назви осей.

- **axes** — логічний аргумент для побудови осей.

- **plot** — логічний аргумент, що відповідає за побудову гістограми.

Якщо **plot** = **TRUE**, то будується гістограма. В іншому випадку функція **hist()** повертає список з інтервалами розбиття і елементами, що попали у відповідні інтервали.

- **labels** — логічний або символьний аргумент.

Якщо **labels** = **TRUE,** то над кожним стовпчиком гістограми виводиться число (або доля) елементів вибірки, що попали в цей інтервал розбиття. Якщо **labels** — символьний аргумент (вектор), то над стовпчиками виводяться елементи цього символьного вектору.

Елементи списку, який повертає функція **hist(x,plot=F)**:

- **breaks** — межі побудованих границь інтервалів.

- **counts** — кількість елементів вибірки, що попали у відповідні інтервали.

- **intensities** — відносні частоти.

- **mids** — середини інтервалів розбиття.

- **xname** — ім’я вектору вибірки.

- **equidist** — чи рівні довжини інтервалів.

attr(,"class") — вказує клас об’єкта, що виводиться.

Приклади:

```
> X <- rnorm(n = 100, mean = 15, sd = 5) # модулюємо нормальну вибірку з
                                         # середнім 15, середнім квадратичним відхиленням 5
> hist(X) # автоматичне розбиття на інтервали і «без кольору»
```

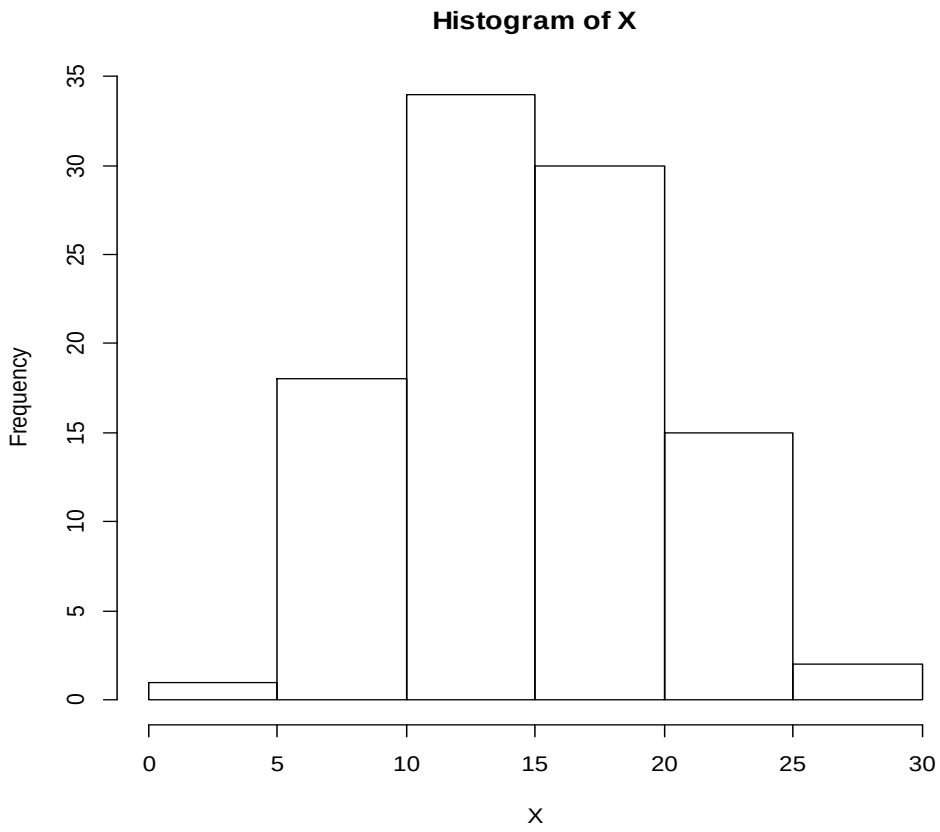

Рис.2 (Малюнок згенеровано **RStudio**)

```
> hist(X, breaks = 20, col = "green") # двадцять інтервалів зеленого кольору
```

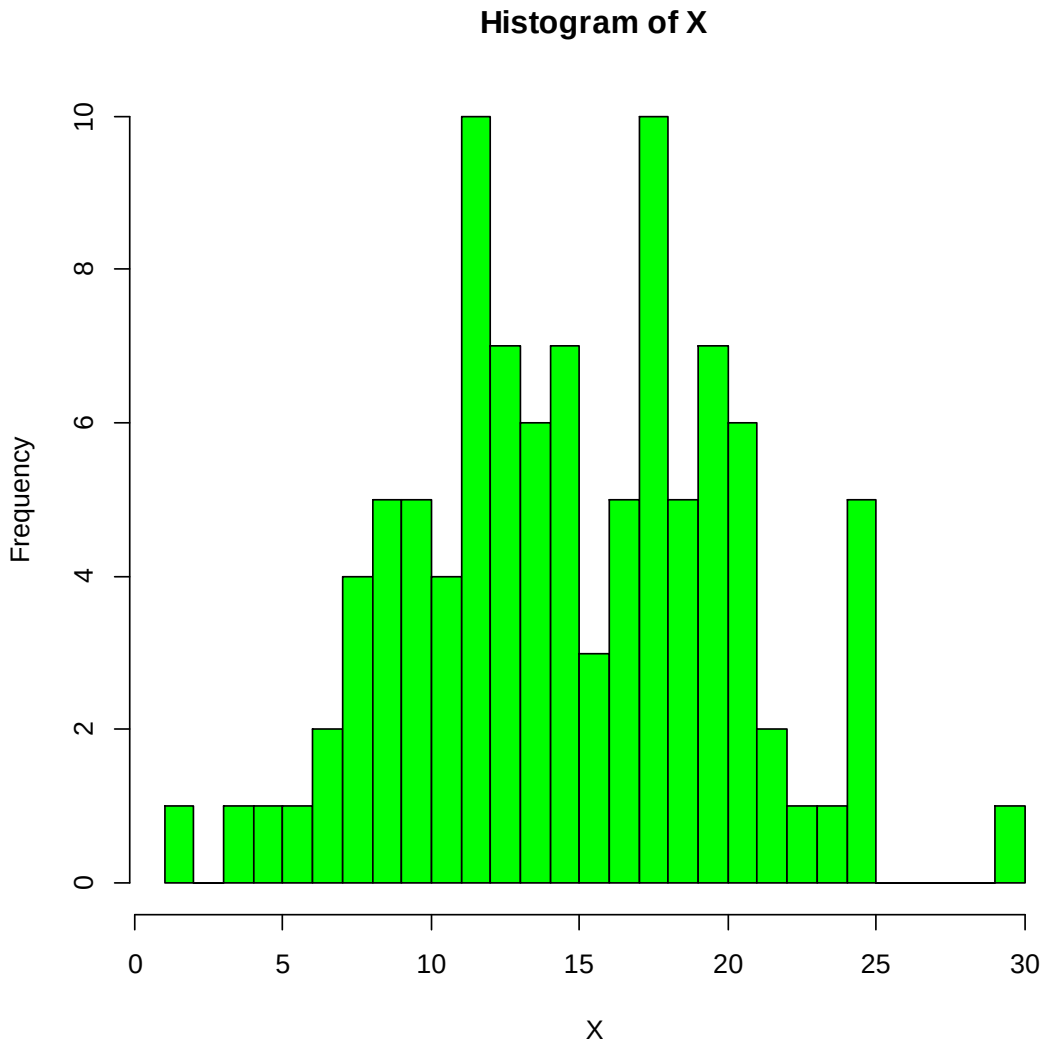

Рис.3 (Малюнок згенеровано **RStudio**)

Якщо параметр **freq = FALSE,** то на графіку відображаються відносні частоти.

Функція **plot(density(x))** зображує оцінку щільності розподілу.

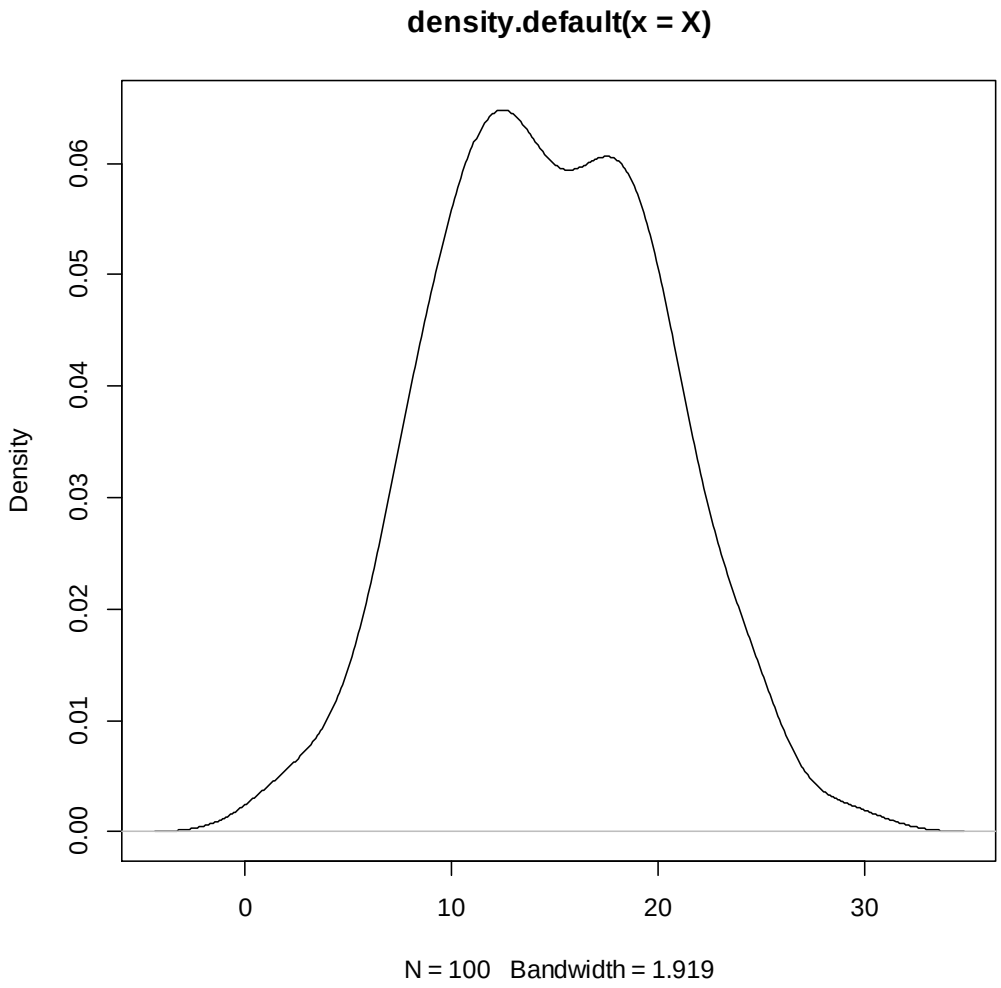

Рис. 4. (Малюнок згенеровано **RStudio**)

Цілком логічно сумістити графік гістограми з графіком щільності:

Приклади:

```
>hist(X, breaks = 20, freq = FALSE, col= "lightblue",
  xlab = " Змінна X",
  ylab = "Щільність",
  main = "Гістограма + графік щільності")
>lines(density(X), col= "red", lwd = 2)
>lines(density(X, bw = 0.8), col= "blue", lwd = 2)
```

Увага! Параметр **bw = 0.8** – параметр вікна ядерної оцінки щільності (Рис.5). Фактично, на малюнку дві оцінки щільності!

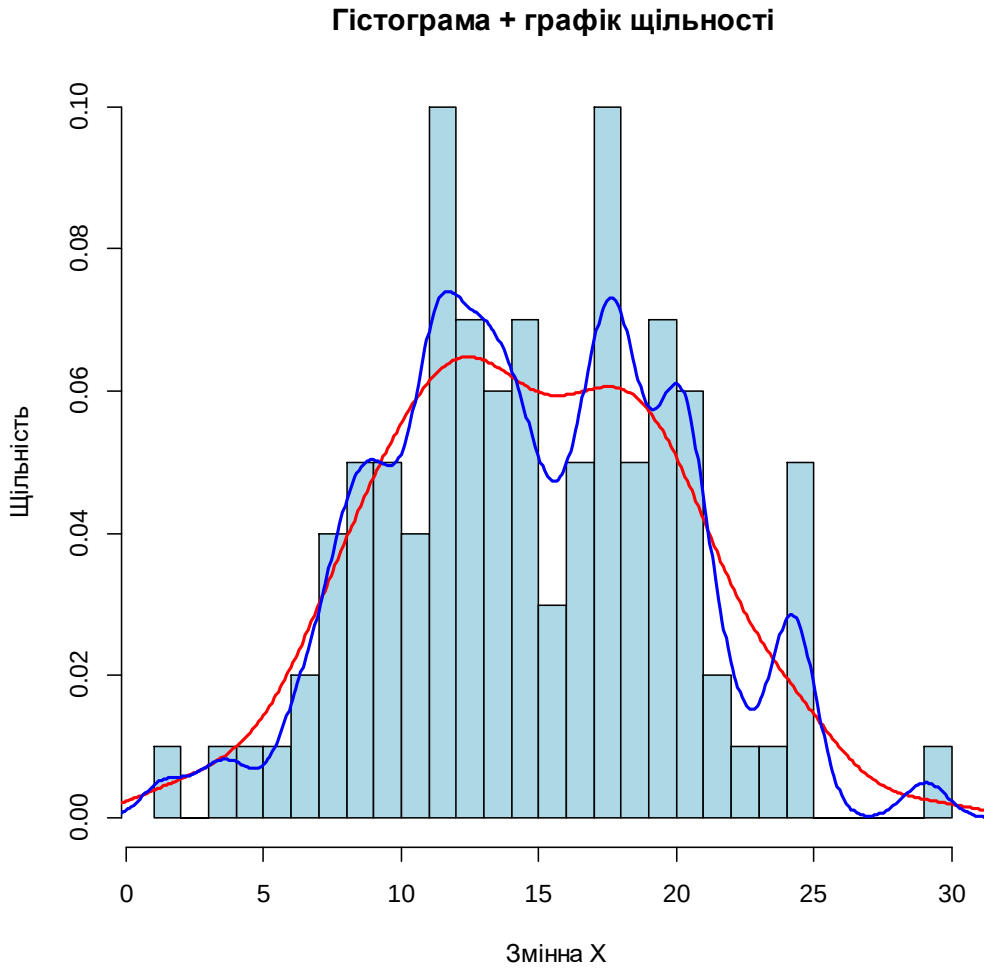

Рис.5. (Малюнок згенеровано **RStudio**)

III. «Скриня з вусами» (box-whisker plots)

Діаграма розмаху – «**скриня з вусами**» - задає візуальну статистичну характеристику генеральної сукупності, основні графічні характеристики (Рис.6):

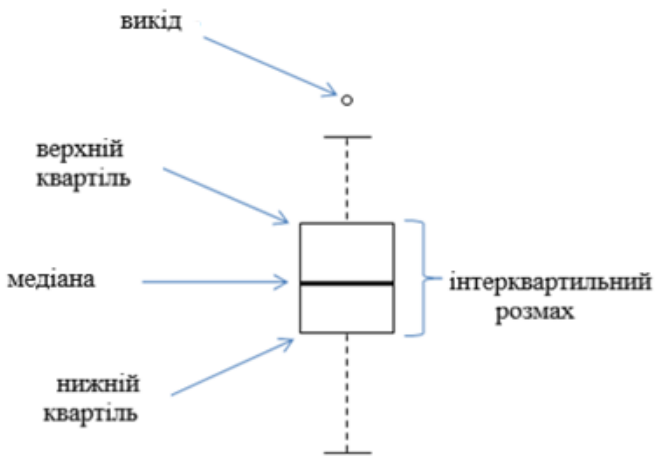

Рис.6 (Малюнок згенеровано **RStudio**)

Задається командою **boxplot(x,..).**

Функція **boxplot()** будує діаграму «скриня з вусами».

Аргументами функції є:

- **x** —або числовий вектор, або список, елементом якого є числові вектори.

- **range** — числовий аргумент, який визначає, наскільки далеко від основної частини діаграми («скрині») знаходяться «вуса» (вертикальні лінії, що відповідають першому і третьому квартілю).

Якщо **range** = 0, то розміщення «вусів» визначається мінімальним і максимальним елементом вибірки.

- **width** — числовий аргумент, що задає ширину коробки діаграми.

Якщо будується декілька «скринь з вусами», то **width** — числовий вектор.

- **varwidth** — логічний аргумент.

Якщо **varwidth = TRUE**, то ширина скрині пропорційна квадратному кореню з об’єму вибірки.

- **notch** — логічний аргумент.

Якщо **notch = TRUE**, то по боках діаграм вирізаються виїмки. Якщо вони не перетинаються (знаходяться на різних рівнях), то можна стверджувати про не співпадіння медіан.

- **outline** — логічний аргумент, що вказує, чи потрібно на графіку малювати викиди.

- **names** — назви окремих діаграм (або символьний аргумент, або типу **expression**).

-**plot** — логічний аргумент , що вказує, чи будується діаграма, чи виводяться результати обробки вибірки).

- **border** — колір меж коробки, «вусів», викидів.

- **col** — колір скрині.

- **log** — символьний аргумент, що показує, чи потрібно осі перетворювати в логарифмічні.

- **pars** — список аргументів, що відповідають за масштаб скрині (**boxwex**), масштаб «вусів» (**staplewex**) і викидів (**outwex**).

- **horizontal** — логічний аргумент, що вказує чи горизонтальні, чи вертикальні будуються діаграми .

- **add** — логічний аргумент, що вказує, чи додавати діаграму до вже існуючих.

- **at** — числовий вектор, що задає порядок побудови діаграм на одному графіку.

Приклади:

```
> boxplot(X,
+ ylab = "Нормальні числа, mean = 15, sd = 5 ",
+ main = "Скриня с вусами",
+ col = "green")
```

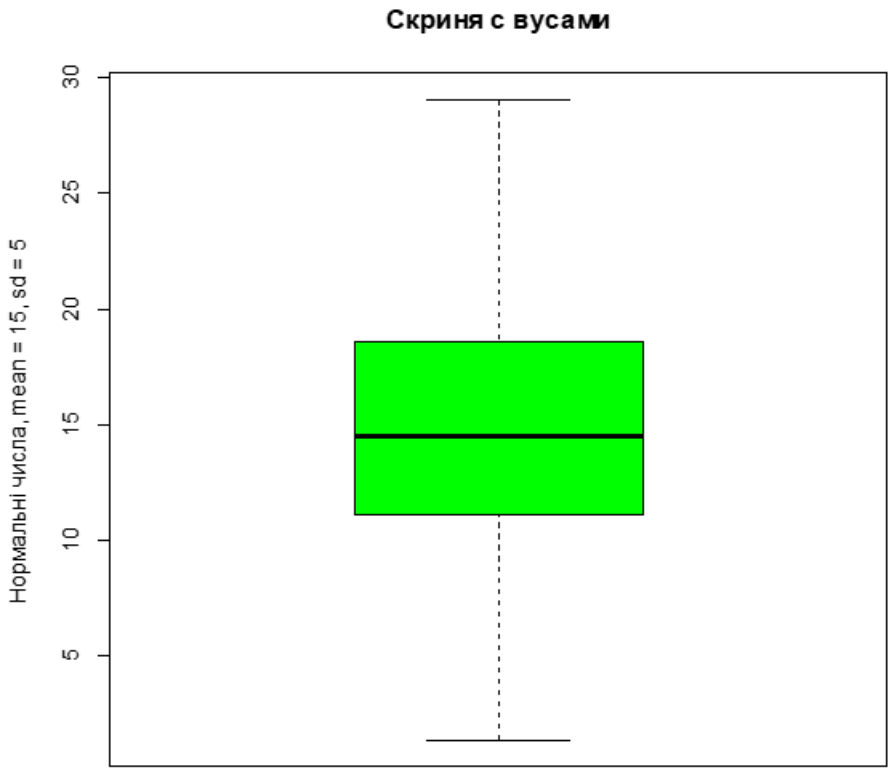

Рис.7 (Малюнок згенеровано **RStudio**)

Нижче (рис.8) для змінних **Ozone, Wind, Temp** з відомого набору даних **airquality** побудовано на одному графіку три «**скрині з вусами**»:

Приклади:

```
> new<-subset(airquality,select=c(Ozone,Wind,Temp))
> boxplot(new,notch=T,col=23,border="green",horizontal = T)
```

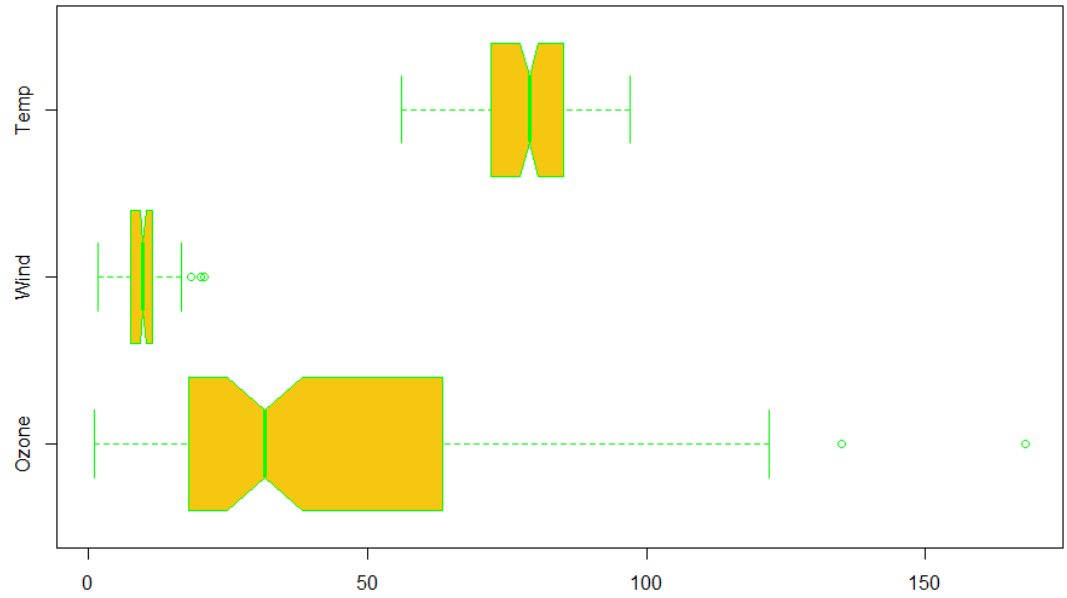

Рис.8 (Малюнок згенеровано **RStudio**)

IV. P-P і Q-Q діаграми

Гістограма є основним і досить зручним статистичним методом дослідження типу і форми розподілу вибірки. Але на її форму дуже впливають співвідношення розмаху вибірки й кількість інтервалів розбиття. При неправильному їхньому виборі можна отримати хибний висновок щодо розподілу. Тому разом з гістограмами використовуються й інші прийоми графічної перевірки того, як емпіричний розподіл даних узгоджується з теоретичним, а саме, **P-P** та **Q-Q** діаграми. Ці діаграми побудовані на порівнянні емпіричної функції розподілу або емпіричних квантілів з теоретичними.

Якщо $X = (x_1, x_2, ..., x_n)$ — вибірка з генеральної сукупності з функцією розподілу $F(x)$, а $F_n(x)$ - емпірична функція розподілу, тоді множину точок з координатами $\left(F(x_i), F_n(x_i)\right), i = \overline{1,n}$ називають **P-P** діаграмою.

Приклади:

```
>       set.seed(3)
>       n<-100
>       x<-rnorm(n) # генеруємо стандартну нормальну вибірку
>       y<-rnorm(n,sd=3) # генеруємо нормальну вибірку з σ = 3
>       # будуємо P-P для x
>        plot(pnorm(sort(x)),(1:n)/n,asp=1,
```

```
+              ylab="Empirical P",
+              xlab="Theoretical P")
>       # пряма  y=x
>         abline(0,1,col=2)
#  P-P діаграма для стандартного розподілу
>         plot(pnorm(sort(y)),(1:n)/n,
+              ylab="Empirical P",
+              xlab="Theoretical P")
>       abline(0,1,col=2)
```

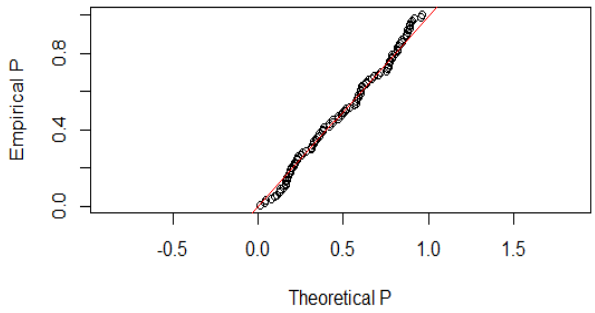

Рис.8. Стандартний нормальний розподіл

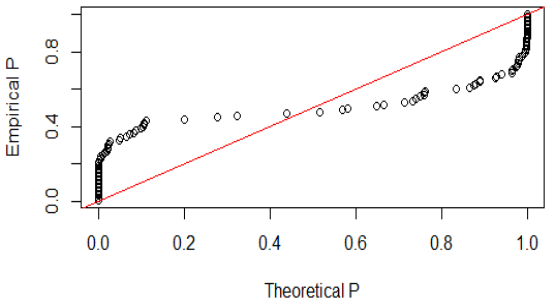

Рис.9. Нормальний розподіл $\sigma = 3$

(Малюнки згенеровано **RStudio**)

В **Q-Q** діаграмі ймовірності замінюють відповідними квантілями. Для нормального розподілу **Q-Q** діаграму у **R** можна побудувати, використовуючи функції **qqnorm()** та **qqline()** (Рис.10):

Приклади:

```
> x<-rnorm(200,mean=1,sd=0.5)
>         qqnorm(x)
>         qqline(x)
```

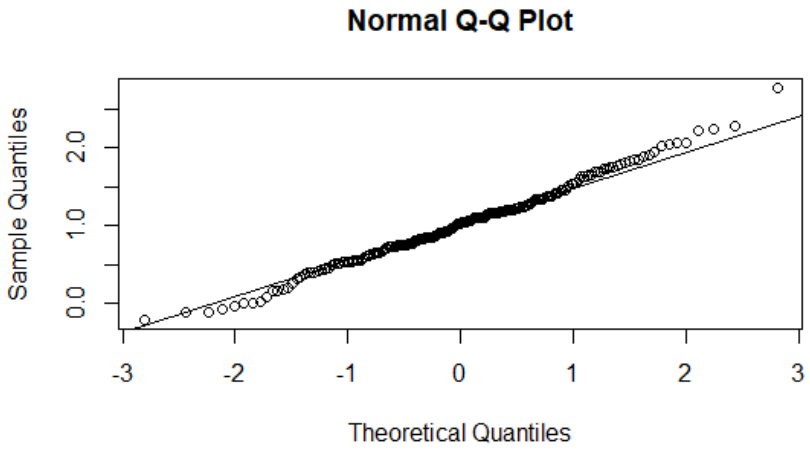

Рис.10 (Малюнок згенеровано **RStudio**)