Course: R Language in Computational Probability and Statistics

Lecture 13: Graphical methods of statistics

Lecturer: Nataliia Kruhlova

Гістограми

Ми вже коротко розглядали функцію **hist()**. Тепер опишемо її більш детально. Функція **hist()** [1] призначена для побудови гістограм. Має наступні параметри:

• **x** — числовий вектор (вибірка).

• **breaks** — параметр, який задає розбиття вибірки на інтервали.

Якщо задається числове значення для цього параметру, то воно визначає кількість інтервалів розбиття. Якщо задається числовий вектор, то він задає межі інтервалів розбиття. Цей параметр може приймати і символьне значення, яке буде визначати алгоритм розбиття на інтервали.

• **freq** і **probability** — два логічних параметри. Рекомендується використовувати тільки один з цих параметрів при побудові, оскільки вони є альтернативними.

Якщо ці параметри не задавати у функції, то будується гістограма частот. Якщо потрібно зобразити гістограму відносних частот, то задають **freq = FALSE** або **probability = TRUE**.

• **include.lowest** — логічний параметр, який використовується в тому випадку, коли **breaks** — числовий вектор. Якщо **include.lowest= TRUE**, то мінімальний елемент вибірки входить в перший інтервал в якості лівої границі.

• **right** — логічний параметр, який визначає вид інтервалів розбиття (якщо **right = TRUE,** маємо відкриті зліва інтервали).

• **density** і **angle** — числові параметри, що відповідають за штрихування стовпчиків гістограми.

• **col** —аргумент, що задає колір заливки чи штриховки стовпчиків гістограми.

• **border** — аргумент, який визначає колір меж стовпчиків гістограми.

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

Якщо **plot = TRUE**, то будується гістограма. В іншому випадку функція **hist()** повертає список з певною статистичною інформацією.

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

Якщо **labels = TRUE**, то над кожним стовпчиком гістограми виводиться кількість елементів вибірки, які попадають в інтервал розбиття.

Інші параметри функції **hist()** співпадають із стандартними параметрами графічних функцій.

Нам будуть потрібні наступні поля списку **hist(x,plot=F)**:

• **breaks** — числовий вектор з межами інтервалів.

• **counts** — частоти попадання елементів вибірки у відповідні інтервали.

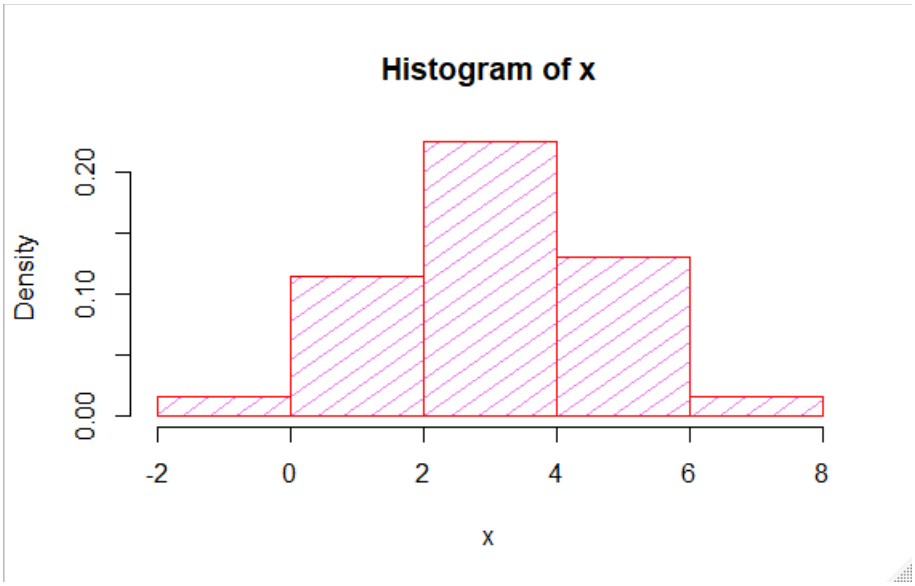
• **density** – містить оцінені значення щільності розподілу.

• **mids** — середини інтервалів розбиття.

Приклад. 1) Згенеруйте нормальну вибірку, для неї побудуйте гістограму відносних частот, стовпчики якої будуть заштрихованими під кутом 35 градусів, інтервалів розбиття буде 5.

2) Використайте функцію **hist(x,plot=F)** для знаходження середнього інтервальної вибірки з 10 рівними інтервалами розбиття.

```
> set.seed(0)
> x<-rnorm(100,3,2)
> hist(x,breaks = 5,freq=F,col="violet",angle=35,density = 10,border="red")
```

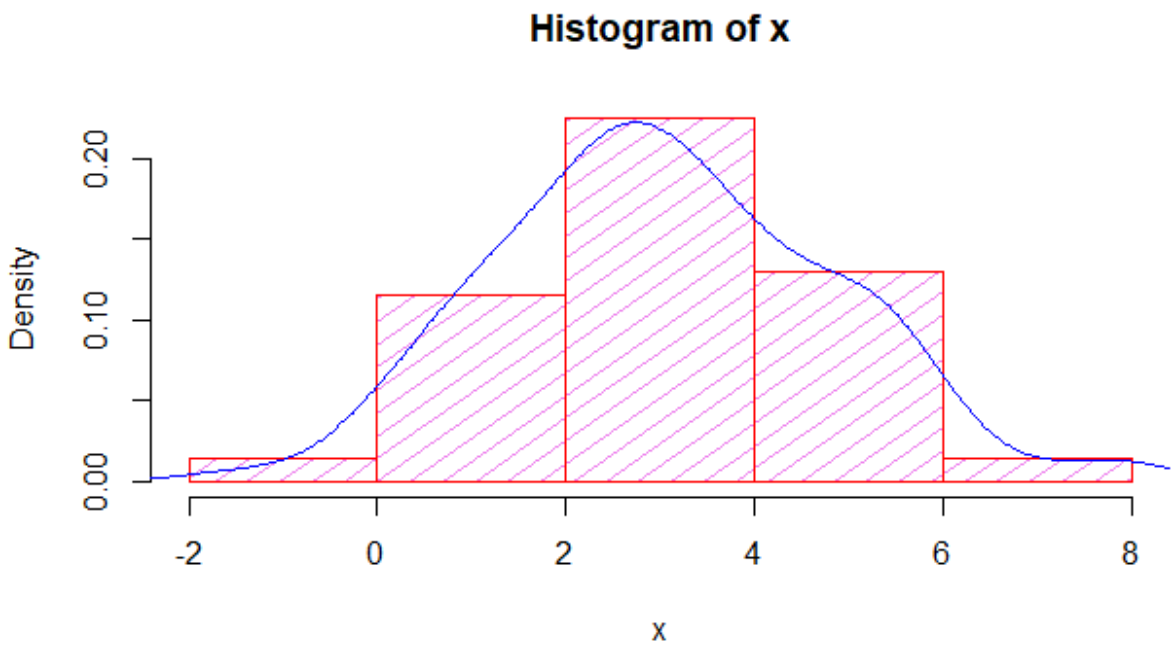


```
> y<-hist(x,breaks=10,plot=F)
> y
$breaks
 [1] -2 -1  0  1  2  3  4  5  6  7  8

$counts
 [1]  1  2  8 15 26 19 14 12  1  2

$density
 [1] 0.01 0.02 0.08 0.15 0.26 0.19 0.14 0.12 0.01 0.02

$mids
 [1] -1.5 -0.5  0.5  1.5  2.5  3.5  4.5  5.5  6.5  7.5

$xname
[1] "x"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
> mean_int<-sum(y$mids*y$counts)/length(x)
> mean_int
[1] 3.06
```

Для визначення форми розподілу змінної використовуються графіки оцінки щільності розподілу. Створюються за допомогою команди **density(x)**.

Приклад. До гістограми з попереднього прикладу додайте оцінку щільності розподілу.

```
> lines(density(x),col="blue")
```



Порівняльні діаграми

Порівняльні діаграми створюються для демонстрації і дослідження залежності змінних від деякого фактору. Кожна діаграма показує залежність змінних від окремого рівня фактору. Дані розбиваються на окремі категорії (наприклад, відповідно до рівнів якогось чинника) і для кожної з них будується свій графік (так звана панель) певного типу. Всі ці графіки потім поєднуються на одному малюнку, що істотно полегшує виявлення статистичних закономірностей та структур у даних.

Порівняльні діаграми будуються функцією **coplot(formula, data, ...).** Параметри функції:

• **formula** задає тип порівняльних діаграм (однофакторні чи двофакторні);

• **data** –дані;

• **panel** - опція, що дозволяє задати тип і налаштувати зовнішній вигляд окремих панелей графіка; за замовчуванням ці панелі є діаграмами розсіювання;

• **rows** і **columns** – задають кількість рядків і стовпчиків матриці панелі графіка;

• **show.given** – логічне аргумент (або вектор із двох логічних значень у випадку двох категоріальних змінних, за якими розбиваються дані), що дозволяє включати (TRUE) або відключати (FALSE) зображення "вивіски" графіка;

• **number** – кількість інтервалів, на які розбиваються змінні **a** та **b** у випадках, якщо ці змінні є чинниками;

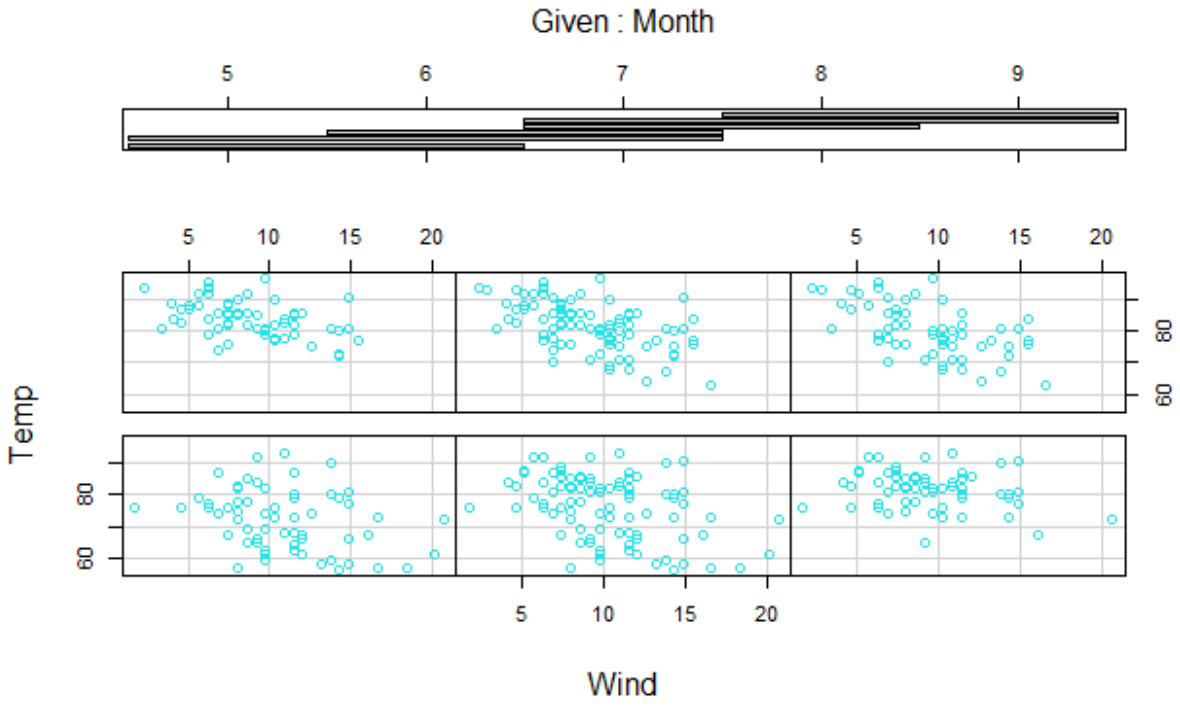
• **overlap** – число (< 1), що визначає область перекриття між даними, які групуються відповідно до рівня кількісних змінних **a** і **b**; якщо **overlap < 0**, відповідна частка спостережень на "перетинах" груп не буде зображуватись.

Формула виду **y~x | a** показує, що графіки залежності від **x** повинні бути побудовані для кожного рівня змінної **a**. У свою чергу, формула виду **y~x| a * b** показує, що графіки залежності від **x** повинні бути одночасно побудовані для кожного рівня як змінної **a**, так і змінної **b**. Усі три чи чотири змінних можуть бути як кількісними, так і якісними (факторами). Якщо **x** або **y** є факторами, то їх рівні будуть автоматично перетворені на чисельні значення (за допомогою функції **as.numeric()**).

Крім перерахованих аргументів, функція coplot() має також такі стандартні графічні параметри, як **col**, **pch**, **xlim**, **ylim** та інш.

Приклад. Розгляньте набір даних **airquality**. Побудуйте порівняльну діаграму залежності змінної **Temp** від **Wind** в залежності від місяця.

```
> coplot(Temp~Wind|Month,data=airquality,col=5)
```



Діаграма «скриня з вусами»

Вже відома нам функція **boxplot()** будує діаграму «скриня з вусами».

Аргументами функції є:

• **x** —або числовий вектор, або список, полями якого є числові вектори.

• **range** — числовий аргумент, який визначає розміщення «вусів» відносно «скрині».

Якщо **range = 0**, то «вуса» задаються мінімумом і максимумом вибірки.

• **width** — числовий параметр, що визначає ширину «скрині».

Цей параметр може бути числовим вектором, якщо **x** - список.

• **varwidth** — логічний аргумент, який визначає, чи ширина «скрині» буде пропорційна квадратному кореню з об'єму вибірки.

• **notch** — логічний аргумент, який вказує, чи потрібно робити виїмки у «скрині».

• **outline** — логічний аргумент, який визначає побудову викидів на діаграмі.

• **plot** — логічний аргумент , що відповідає за побудову діаграми.

• **border** — задає колір меж діаграми і викидів.

• **col** — визначає колір «скрині».

• **pars** — список параметрів, що відповідають за масштаб елементів діаграми.

• **horizontal** — логічний параметр, що положення діаграми.

• **add** — логічний параметр, що дозволяє додавати нову діаграму до існуючих.

Горизонтальна лінія всередині прямокутника - це медіана вибірки, верхня і нижня межі прямокутника — це 0.75 и 0.25 квантилі вибірки. Верхня і нижня вертикальні лінії або відповідають максимальному і мінімальному значенню вибірки, або це відступи на 1.5IQR вверх і вниз від медіани. Точки, що лежать за цими лініями відповідають викидам.

Діаграма «скриня з вусами» показує вид розподілу елементів вибірки і симетричність. Також використовується для виявлення помилкових даних (викидів).

Функція **boxplot(x,plot=F)** використовується для визначення наступної статистичної інформації:

• **stats** — матриця, число стовпчиків якої відповідає числу вибірок, до яких застосовувалась функція **boxplot**. Елементами кожного стовпчика є: мінімальне значення вибірки, перший квартиль, медіана, третій квартиль і максимальне значення вибірки.

• **n** — об’єм вибірки.

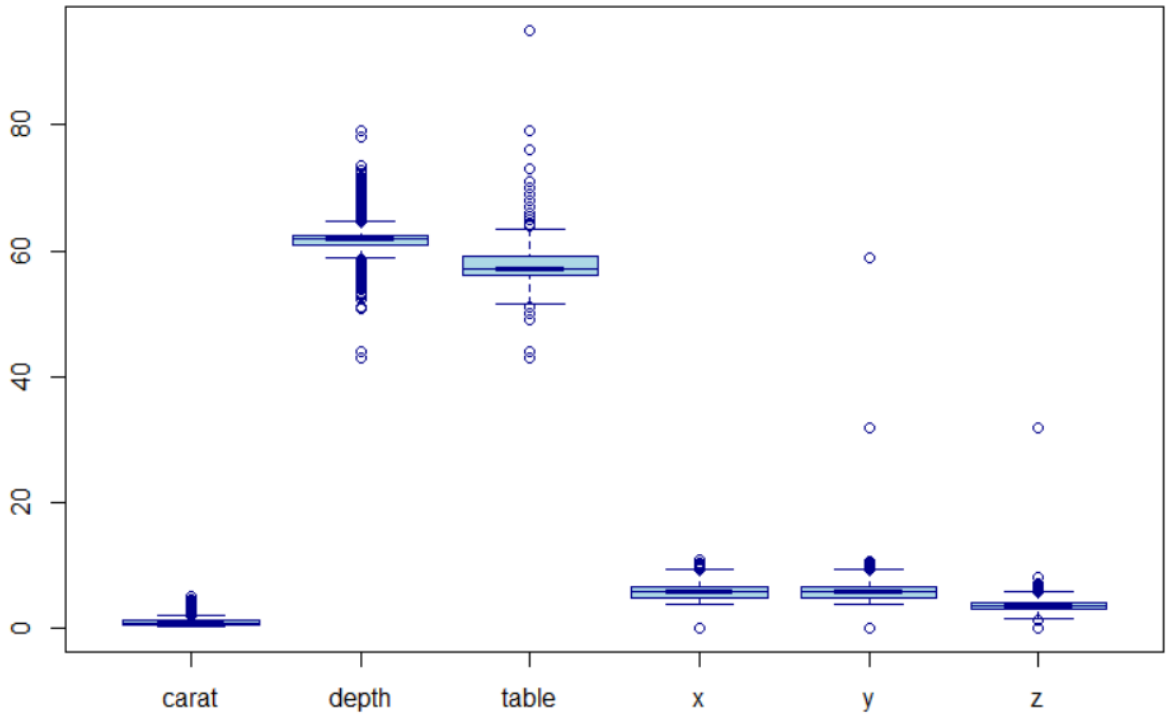
• **out** — викиди.

• **group** — вектор такої ж довжини, як і **out**, вказує до якої вибірки належать викиди.

• **names** — імена вибірок.

Приклад. Для всіх числових змінних, крім змінної **price**, з набору даних **diamonds** з пакету **ggplot2** побудуйте діаграми «скриня з вусами». Потім без побудови знайдіть кількість викидів для кожної змінної.

```
> library(ggplot2)
> data<-diamonds[,-c(2:4,7)]
> boxplot(data,notch = T,col="lightblue",border="darkblue")
> y<-boxplot(data,plot=F)
> tapply(y$out,y$group,length)
   1    2    3    4    5    6
1889 2545  605   32   29   49
```



Q-Q і P-P діаграми

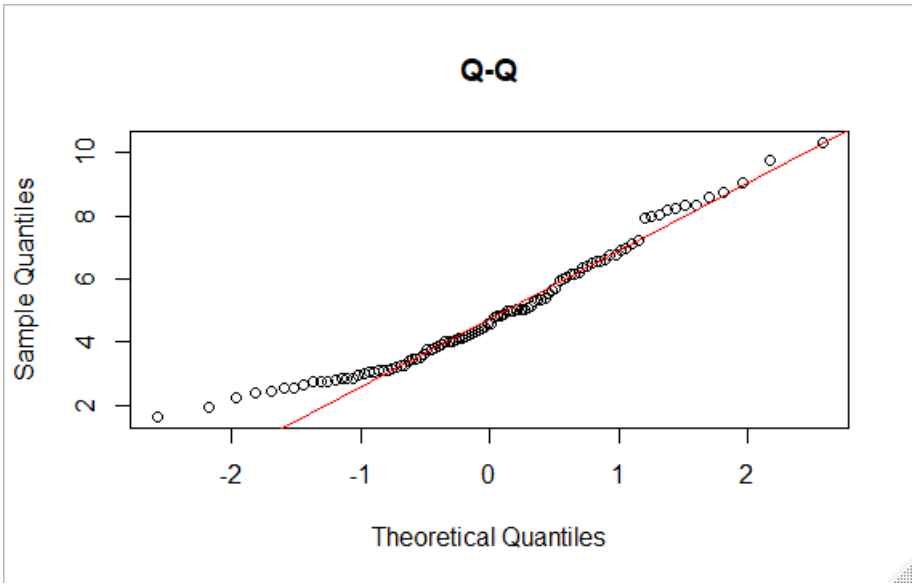
Ці дві діаграми використовуються для перевірки, чи вибірки узгоджуються з певним законом розподілу. Q-Q діаграма порівнює емпіричні і теоретичні квантилі, а P-P діаграма порівнює емпіричну і теоретичну функції розподілу. Також ці графіки можна використовувати для візуальної оцінки того, чи співпадають два розподіли даних.

Функція **qqnorm()** порівнює емпіричні квантилі з теоретичними для нормального розподілу.

Функція **qqline()** проводить лінію на нормальному графіку «квантиль-квантиль» через перший і третій квартилі.

Приклад. Побудуйте Q-Q діаграму для нормального розподілу.

```
> x<-rnorm(100,5,2)
> qqnorm(x,main="Q-Q")
> qqline(x,col="red")
```



Функція **qqplot()** призначена для побудови Q-Q діаграми.

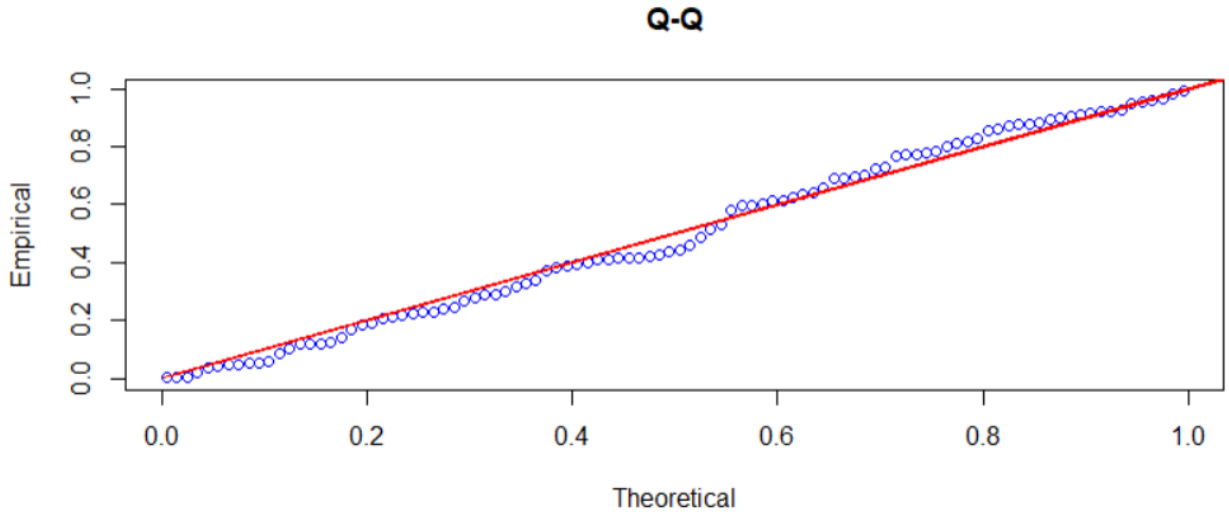
Аргументами є:

• x і y — два числових вектори.

• plot.it — логічний аргумент , що вказує, чи потрібно будувати графік.

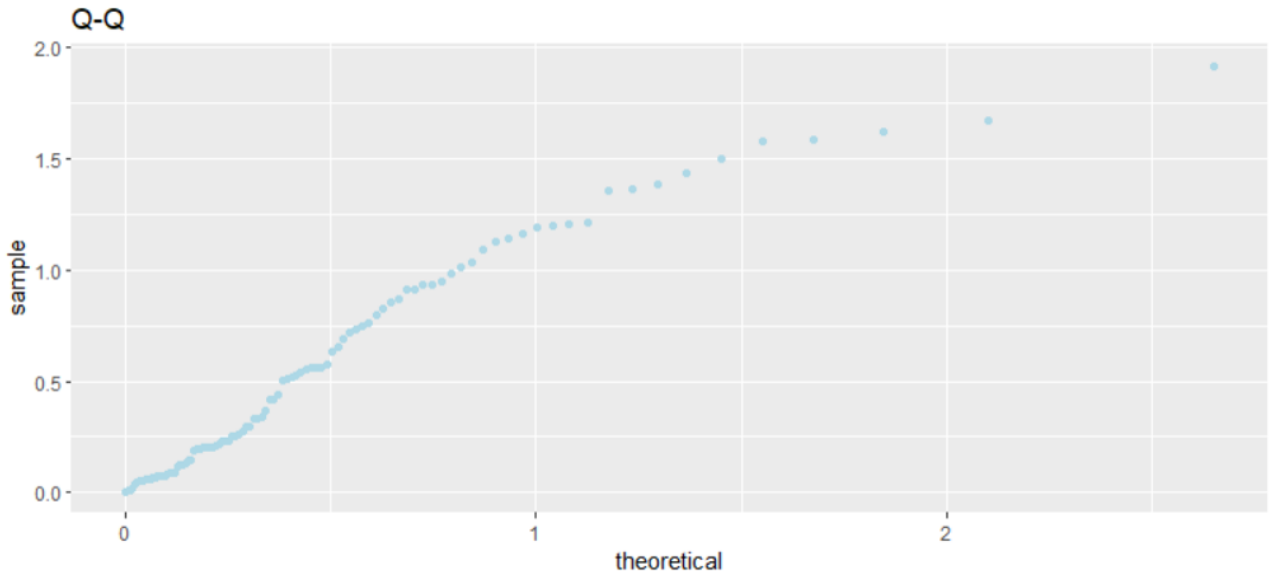
Приклад. Згенеруємо рівномірну вибірку і перевіримо, чи вона узгоджується з рівномірним законом розподілом. Побудуємо Q-Q діаграму.

```
> y<-runif(100)
> qqplot(qunif(ppoints(100)),y,col=»blue»,main=»Q-Q»,
ylab=»Empirical»,xlab=»Theoretical»)
> lines(x,x,col=»red»)
```



Приклад. Згенеруйте вибірку експоненціального розподілу. Побудуйте Q-Q діаграму за допомогою пакету **ggplot2**.

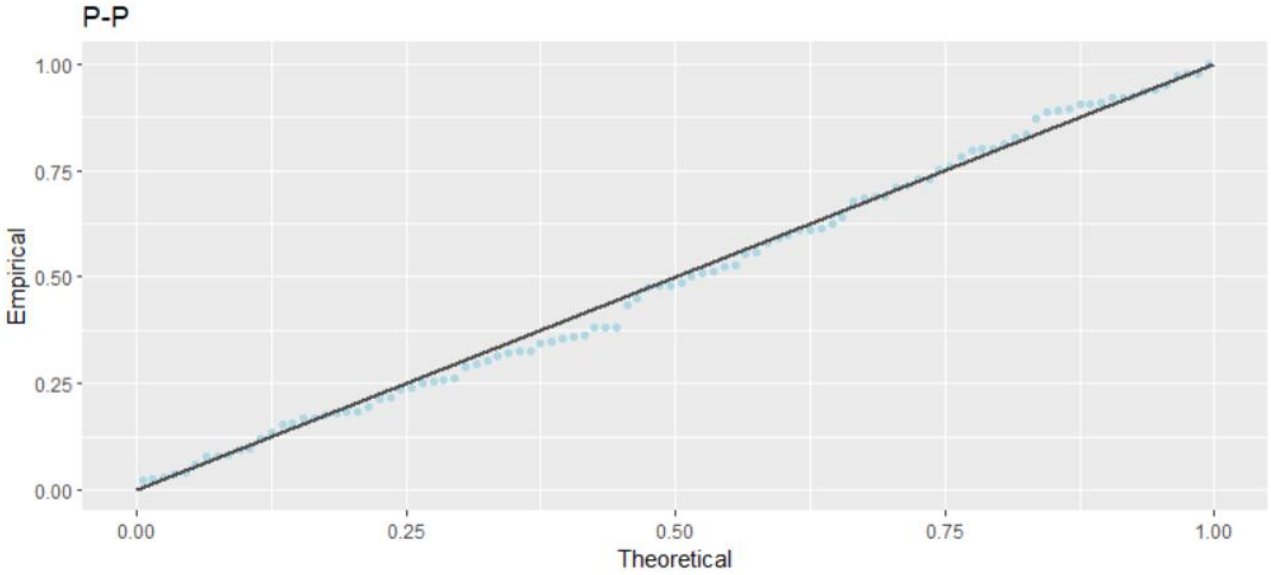
```
> library(ggplot2)
> x<-rexp(100,2)
> ggplot(as.data.frame(x),aes(sample=x))+
stat_qq(distribution=qexp,dparams=2,col="lightblue")+
labs(title="Q-Q")
```



Приклад. Згенеруйте вибірку нормального розподілу. Використайте пакет **qqplotr** [2] для побудови P-P діаграми.

```
> library(qqplotr)
```

```
> x<-rnorm(100)
> ggplot(as.data.frame(x),aes(sample=x))+
stat_pp_point(col="lightblue")+
stat_pp_line()+
labs(title="P-P",x="Theoretical",y="Empirical")
```



Всі малюнки в лекції були згенеровані в RStudio.

Список джерел

1. The R Project for Statistical Computing. https://www.r-project.org
2. An Introduction to qqplotr. https://cran.r-project.org/web/packages/qqplotr/vignettes/introduction.html