Лабораторна робота № 3. Основи вибіркового методу

Гладкий Іван

Мета: *засвоїти основи статистичного оцінювання характеристик випадкової величини на основі вибіркового підходу засобами мови програмування R; набути навичок роботи у середовищі RStudio із застосуванням концепції "грамотного програмування" із застосуванням пакету R Markdown.*

1. Постановка задачі

Випадкова величина X має нормальний закон розподілу: $X \sim F(a, \sigma^2)$, тобто $X \sim N(a, \sigma^2)$, вектор параметрів $\Theta = (a, \sigma^2)$ якого відомий: $(a, \sigma^2) =$ (-1 , 1) . Тобто, $a =$ -1, $\sigma^2 = 1$. Згенерувати дві вибірки випадкової величини X за допомогою відповідного генератору псевдовипадкових чисел: $(x_1, x_2, ..., x_n)$ відповідно обсягів $n = 100$ та $n = 1000$, що мають розподіл $X \sim N($ -1, 1), обчислити і дослідити оцінки параметрів розподілу $\tilde{a} = \tilde{a}(x_1, x_2, ..., x_n) \approx a, \tilde{\sigma} = \tilde{\sigma}(x_1, x_2, ..., x_n) = \sigma$ та інші статистичні характеристики, зробити порівняльний аналіз оцінених характеристик між собою і з теоретичними характеристиками. Для цього необхідно:

1. Побудувати статистичний розподіл у вигляді інтервальної таблиці відносних частот.
2. Побудувати гістограму, теоретичну $f(x, a, \sigma)$ та емпіричну $f^*(x, \tilde{a}, \tilde{\sigma})$ функції щільності на одному графіку.
3. Побудувати графіки теоретичної $F(x, a, \sigma)$ та емпіричної $F^*(x, \tilde{a}, \tilde{\sigma})$ функції розподілу.
4. Побудувати п'ятиквантильний графік (boxplot) "ящик з вусами".
5. Обчислити точкові незміщені і конзістентні оцінки вектору параметрів розподілу (a, σ^2), математичного сподівання $m(x)$, дисперсії $D(x)$, СКВ $\sigma(x)$, центральних теоретичних моментів 3-го μ_3 і 4-ого μ_4 порядків, асиметрії A_s та ексцесу E_k:

- написавши власну користувацьку функцію;
- за допомогою вбудованих засобів R.

6. Дані звести у таблицю:

Таблиця 1: Таблиця 1 – **Теоретичні та емпіричні (вибіркові) числові характеристики випадкової величини**

Назва числової характеристики	теоретичне значення	Вибіркове значення, $n = 100$	Вибіркове значення, $n = 1000$
a	a	$\tilde{a}$	$\tilde{a}$
σ	σ	$\tilde{\sigma}$	$\tilde{\sigma}$
Математичне сподівання	$m(x)$	$\tilde{m}(x)$	$\tilde{m}(x)$

Назва числової характеристики	теоретичне значення	Вибіркове значення, $n = 100$	Вибіркове значення, $n = 1000$
Дисперсія	$D(x)$	$\tilde{D}(x)$	$\tilde{D}(x)$
Виправлена дисперсія		$\tilde{\tilde{D}}(x)$	$\tilde{\tilde{D}}(x)$
СКВ	$\sigma(x)$	$\tilde{\sigma}(x)$	$\tilde{\sigma}(x)$
Виправлене СКВ		$\tilde{\tilde{\sigma}}(x)$	$\tilde{\tilde{\sigma}}(x)$
Центральний момент 3-го порядку	μ_3	$\tilde{\mu}_3$	$\tilde{\mu}_3$
Центральний момент 4-го порядку	μ_4	$\tilde{\mu}_4$	$\tilde{\mu}_4$
Асимерія	A_s	$\tilde{A}_s$	$\tilde{A}_s$
Ексцес	E_k	$\tilde{E}_k$	$\tilde{E}_k$

1. Виконання роботи

Генеруємо вибірку з нормального розподілу з параметрами a = -1, $\sigma = 1$ об'єму $n = 100$:

```
set.seed(0) #
X <- rnorm(n, a, s) #
cat("\n", "          :", "\n", "a = ", a, "\n" , "s = ", s, "\n" , "n = ", n, "\n")
```

```
##
##           :
##  a =  -1
##  s =  1
##  n =  100
```

```
cat("             :", head(X))
```

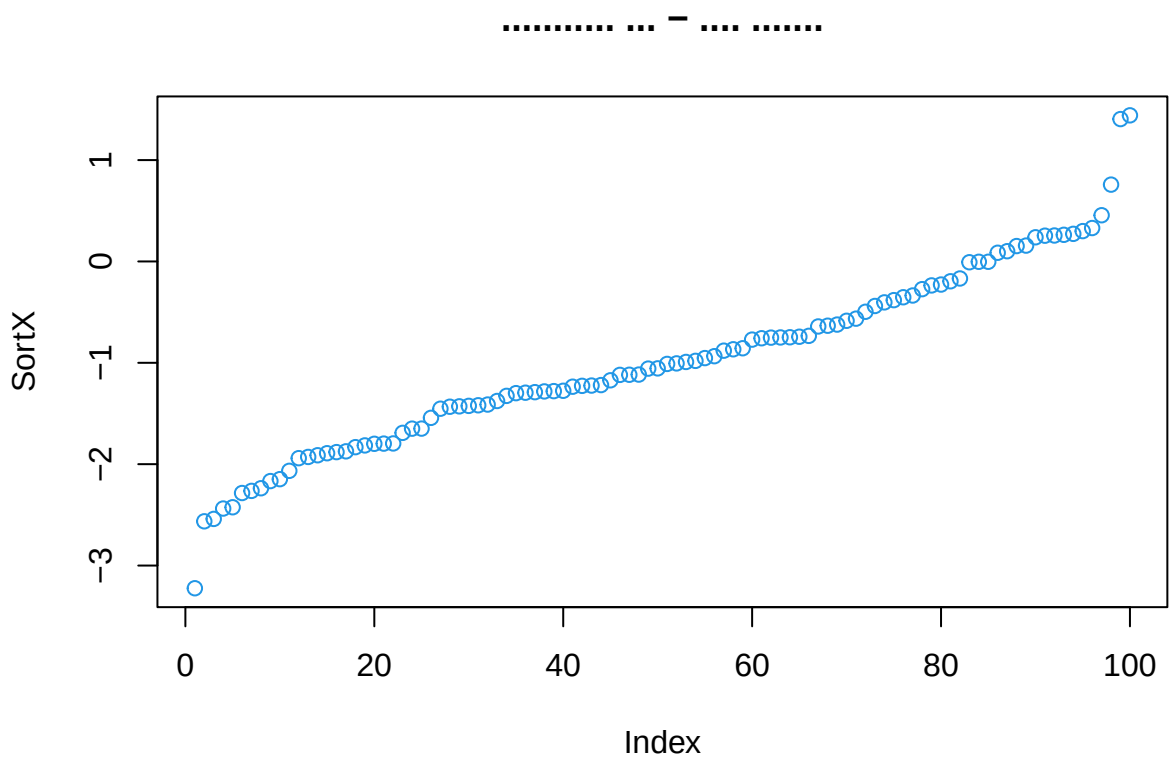
```
##              : 0.2629543 -1.326233 0.3297993 0.2724293 -0.5853586 -2.53995
```

```
cat("              :", tail(X))
```

```
##               : -0.403741 -0.8802824 -1.282174 0.4559884 -0.7709804 -0.003456071
```

Будуємо варіаційний ряд.

```
SortX <- sort(X) #
plot(SortX, col=4)
title("          -        ")
```




```
# dev.off()
```

Будуємо інтервальний статистичний розподіл і гістограму частот.

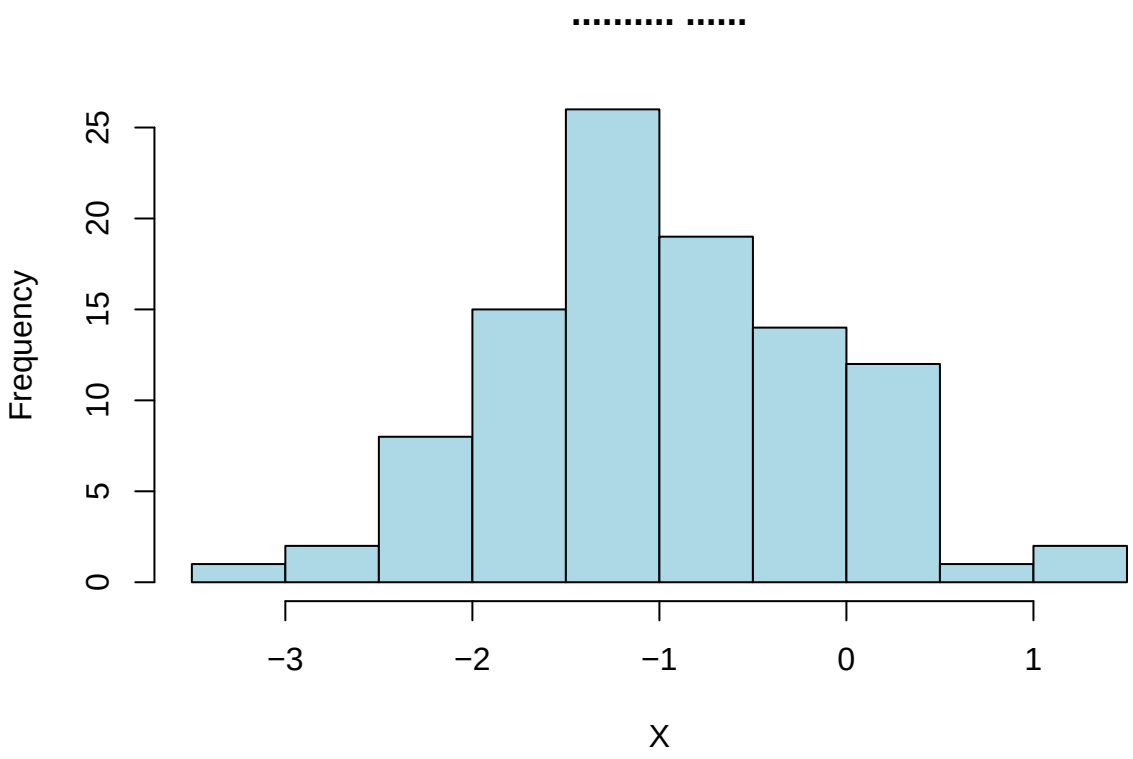
```
1 + 1.332 * log(n)
```

```
## [1] 7.134087
```

```
#          - "Sturges",
table(cut(X, nclass.Sturges(X))) #
```

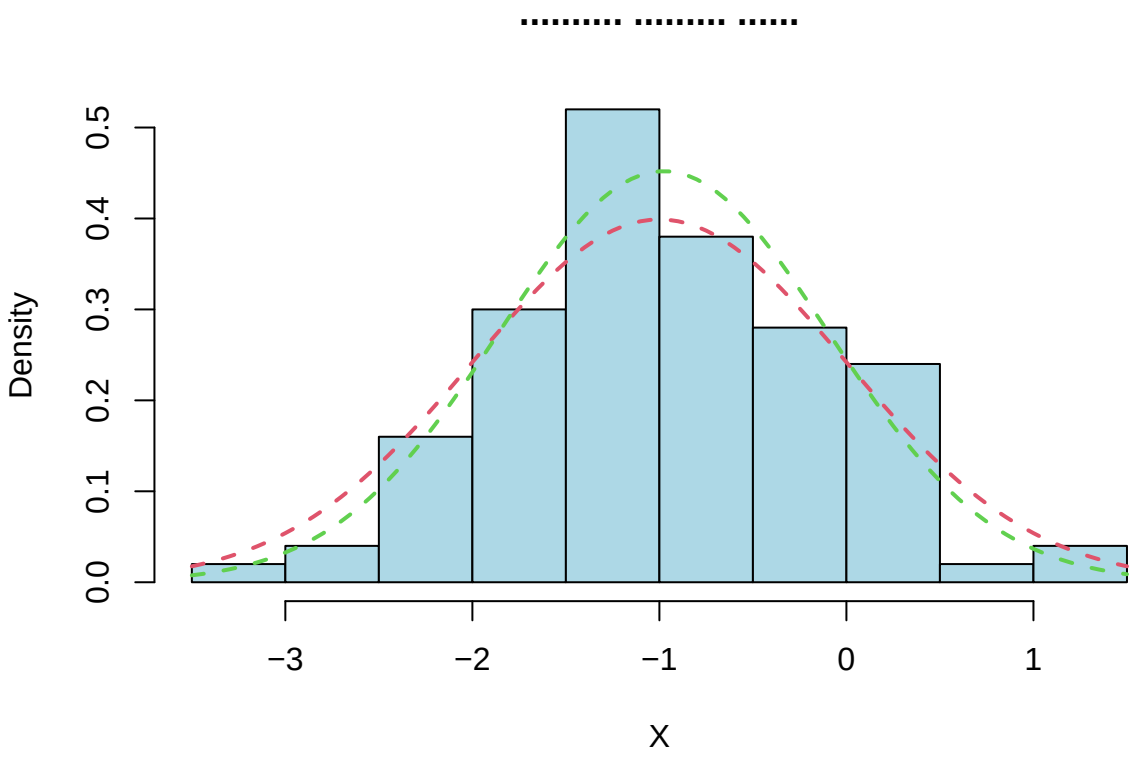
```
##
##   (-3.23,-2.64]   (-2.64,-2.06]   (-2.06,-1.47]  (-1.47,-0.891] (-0.891,-0.308]
##               1              10              15              30              21
##  (-0.308,0.275]   (0.275,0.858]    (0.858,1.45]
##              17               4               2
```

```
hist(X, breaks=nclass.Sturges(X),
     col="Lightblue",
     main="          ") #
```



Гістограма відносних частот.

```
hist(X,
     freq = FALSE,
     col = "Lightblue",
     main = "                ")
curve(dnorm(x, a, s), col = 2, lty = 2, lwd = 2, add = TRUE) #
curve(dnorm(x, mean(X), sd(X)), col = 3, lty = 2, lwd = 2, add = TRUE) #
```



Сумарні статистики.

Вивід сумарних статистик.

```
summary(X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.2239 -1.5694 -1.0330 -0.9773 -0.3746  1.4414
```

Можна так:

Таблица 2: Таблиця 2 – **Числові характеристики вибірки**

Характеристика	Значення
$\tilde{m}(x)$	-0.9773316
$\tilde{\tilde{\sigma}}(x)$	0.8826502
$\tilde{R}$	1.1947698

Функція `describe()` з пакету `Hmisc` надає можливість вивести цілу низку сумарних оцінок характеристик вибірки:

```
describe(X)
```

```
## X
```

```
##        n  missing distinct     Info     Mean  pMedian      Gmd      .05
##      100        0      100        1  -0.9773  -0.9989   0.9995  -2.2916
##      .10      .25      .50      .75      .90      .95
##  -2.0738  -1.5694  -1.0330  -0.3746   0.2399   0.3008
##
## lowest : -3.2239  -2.56378 -2.53995 -2.43759 -2.4251
## highest: 0.329799 0.455988 0.757903 1.40465  1.44136
```

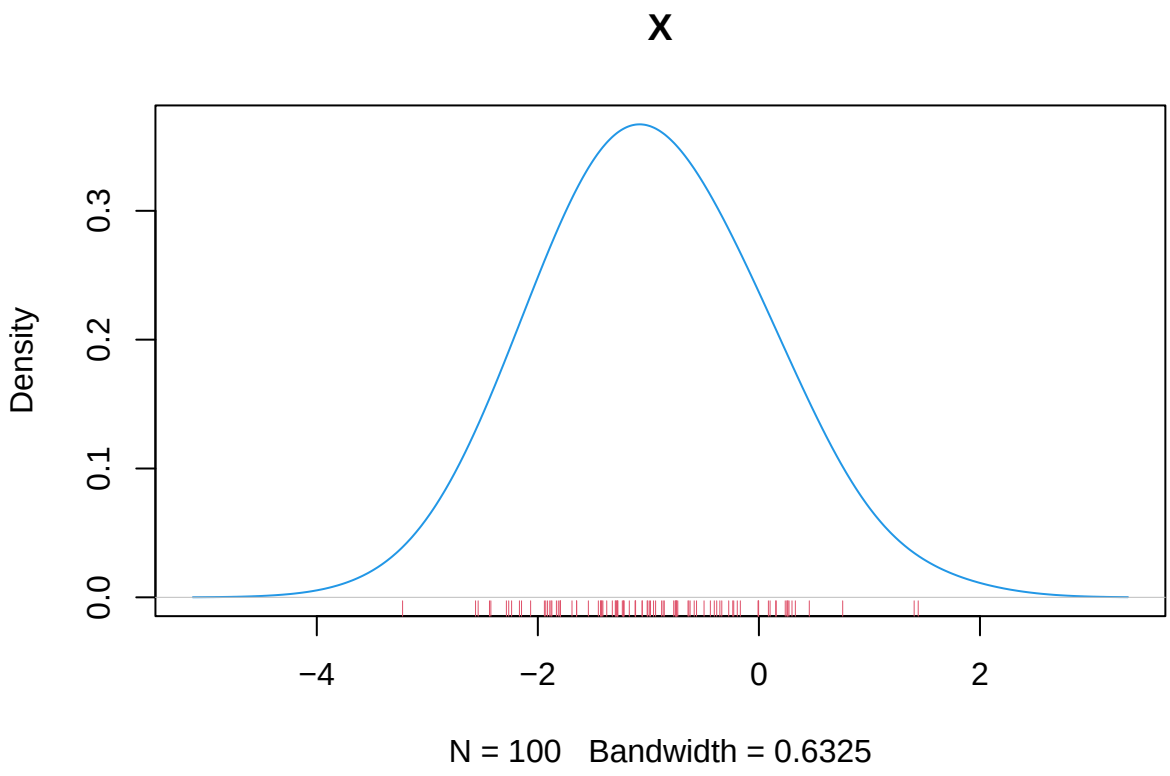
А можна написати власну функцію.

```
my_summary <- function(X, ...) {
  cat("\n", "                          :", "\n",
      "                    , m = ", mean(X), "\n",
      "    ,   = ", median(X) , "\n",
      "            , s^2 = ", var(X), "\n" ,
      "         , s = ", sd(X), "\n" ,
      "    , R = ", max(X) - min(X), "\n" ,
      "              , IQR = ", IQR(X), "\n",
      "    , Ek = ", kurtosis(X), "\n",
      "      , As= ", skewness(X), "\n"
    )
}

my_summary(X,  na.rm=FALSE)
```

```
##
##                         :
##                    , m =  -0.9773316
##      ,   =  -1.032961
##              , s^2 =  0.7790714
##          , s =  0.8826502
##      , R =  4.665265
##                , IQR =  1.19477
##      , Ek =  2.959992
##       , As=  0.2273757
```

Ще один варіант представлення теоретичного і емпіричного розподілів.

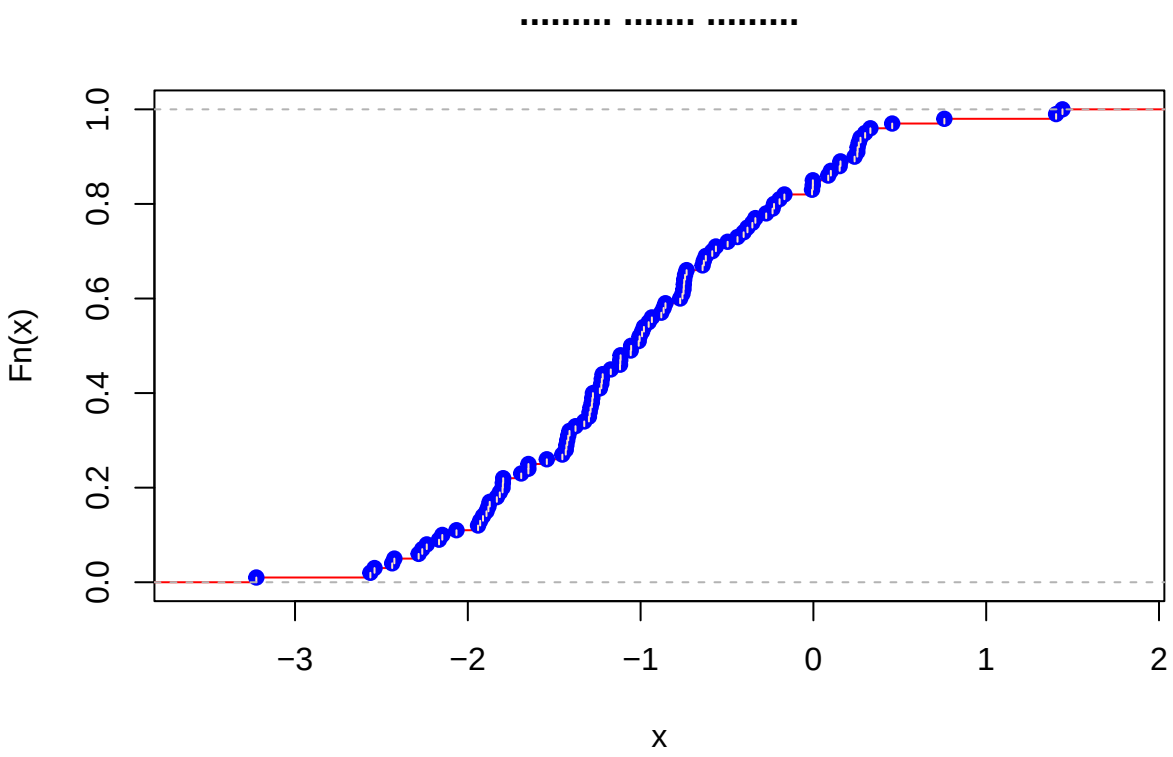
```
plot(density(X, adjust=2), main="X", col=4)
rug(X, col=2) #
```



Емпірична функція росподілу F(x).

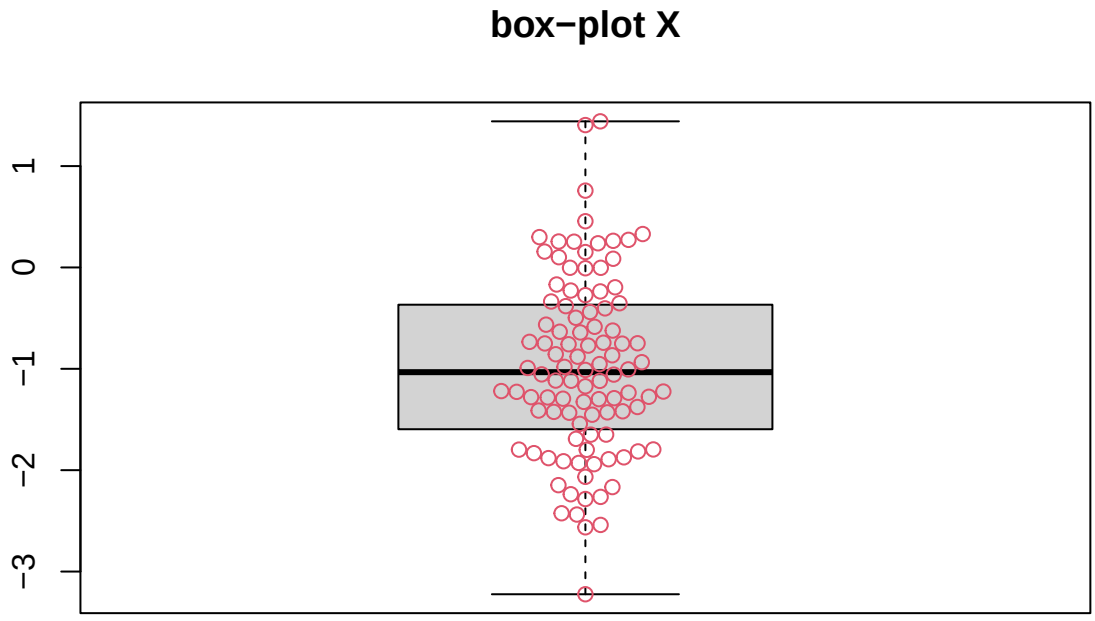
F(x) з числом інтервалів за Стерджессом.

```
# Fn <- ecdf(table(cut(X, nclass.Sturges(X))))

Fn <- ecdf(X)


## luxury plot
plot(Fn,
     verticals = TRUE,
     col.points = "blue",
     col.hor = "red",
     col.vert = "bisque",
     main = "                ")
```



Побудова боксплотів (ящиків з вусами).

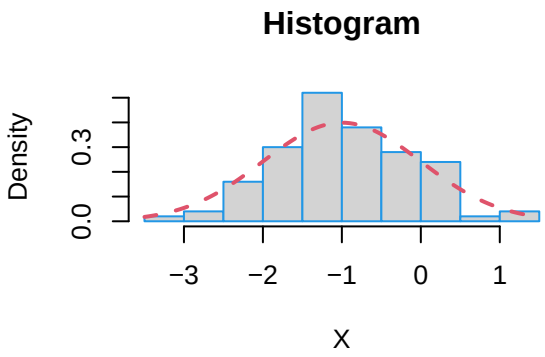
```
#              - "   -   "
boxplot(X, main="box-plot X")  #
beeswarm(X, col=2, add=TRUE) #       "  "       jutter
```

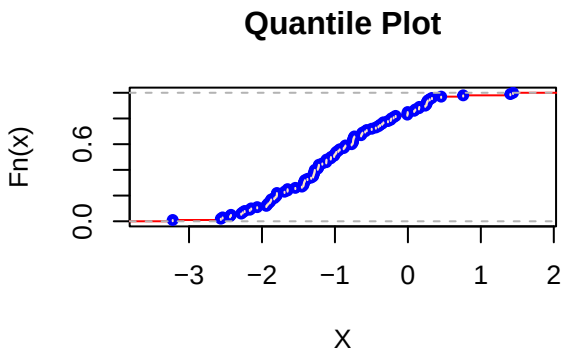


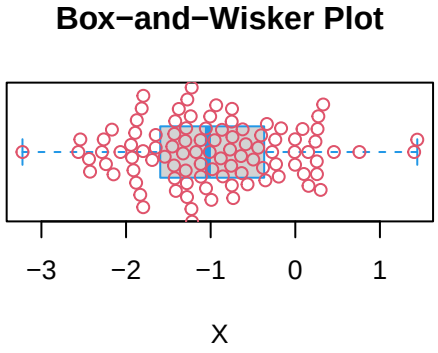
На практиці оптимальним може бути компактний варіант виводу основних графіків вибіркових характеристик. Наприклад, такий.

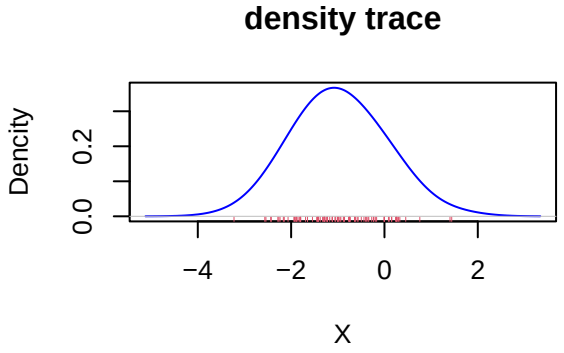
```
op <- par(mfrow = c(2,2))

#
hist(X,
     freq = FALSE,
     col = "Lightgray",
     main="Histogram",
     border=4)

curve(dnorm(x, a, s),
      col = 2,
      lty = 2,
      lwd = 2,
      add = TRUE) #

plot(Fn,
     main = "Quantile Plot",
     verticals = TRUE,
     col.points = "blue",
     col.hor = "red",
     col.vert = "bisque",
     xlab = "X",
     ylab = "Fn(x)")
```

```r
boxplot(X,
        main = "Box-and-Wisker Plot",
        col = "Lightgray",
        border = 4,
        xlab = "X",
        ylab = "",
        horizontal = TRUE)  #

beeswarm(X,
         col = 2,
         add = TRUE,
         horizontal = TRUE) #       " "       jutter

plot(density(X, adjust=2),
     main = "density trace",
     xlab = "X",
     ylab = "Dencity",
     col="blue")

rug(X,
    col=2,
    main="fn(x)") #
```









```r
par(op) #
```

Результати моделювання зводимо у таблицю.

Таблица 3: Таблиця 3 – **Теоретичні та емпіричні (вибіркові) числові характеристики випадкової величини**

Назва числової характеристики	теоретичне значення	Вибіркове значення, $n = 100$	Вибіркове значення, $n = 1000$
a	-1	-0.9773316	$\tilde{a}$
σ	1	0.8826502	$\tilde{\sigma}$
Математичне сподівання	-1	-0.9773316	$\tilde{m}(x)$
Дисперсія	1	0.7790714	$\tilde{D}(x)$
Виправлена дисперсія		0.7869408	$\tilde{\tilde{D}}(x)$
СКВ	1	0.8826502	$\tilde{\sigma}(x)$
Виправлене СКВ		0.8870968	$\tilde{\tilde{\sigma}}(x)$
Центральний момент 3-го порядку	μ_3	$\tilde{\mu}_3$	$\tilde{\mu}_3$
Центральний момент 4-го порядку	μ_4	$\tilde{\mu}_4$	$\tilde{\mu}_4$
Асимерія	A_s	$\tilde{A}_s$	$\tilde{A}_s$
Ексцес	E_k	$\tilde{E}_k$	$\tilde{E}_k$

Контрольні Питання

1. Вибірка — це підмножина даних, обрана з генеральної сукупності для дослідження її властивостей.

2. Вибіркове математичне сподівання оцінюється за допомогою середнього арифметичного. Воно є точковою оцінкою математичного сподівання генеральної сукупності.

3. Дисперсія, дисперсія, середнє квадратичне відхилення, коефіцієнт варіації.

4. Асиметрія, Ексцес.

Висновок

На цьому занятті я засвоїв основи статистичного оцінювання характеристик випадкової величини на основі вибіркового підходу засобами мови програмування R; набув навичок роботи у середовищі RStudio із застосуванням концепції “грамотного програмування” із застосуванням пакету R Markdown