**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**Національний університет «Запорізька політехніка»**

кафедра програмних засобів

**ЗВІТ**

з лабораторної роботи № 2

з дисципліни «Емпіричні методи в інформаційних технологіях» на тему:

" СТАТИСТИЧНА ПЕРЕВІРКА ГІПОТЕЗ"

Варіант 15

Виконала:

студентка групи КНТ-132 Лещинська А.Р

Прийняла:

доцент кафедри Подковаліхіна О.О

2024

**2 ЛАБОРАТОРНА РОБОТА №2**

**СТАТИСТИЧНА ПЕРЕВІРКА ГІПОТЕЗ**

**Мета роботи**

Вивчити методику статистичної перевірки гіпотез. Отримати основні характеристики розподілу ймовірностей випадкової величини та перевірити гіпотезу про закон розподілу вибірки з використанням пакету Statgraphics та мови програмування R.

**Завдання на лабораторну роботу:**

1 Використовуючи рекомендовану літературу та ці методичні вказівки вивчити основні поняття та застосування методики перевірки гіпотез для оцінювання параметрів випадкових величин, роботу статистичного пакету програм Statgrарнісs та мови програмування R, для перевірки статистичних гіпотез розподілу випадкових величин.

2 Вивчити загальні положення теорії статистичної перевірки гіпотез.

3 Згенерувати стовпець даних на основі наступної інформації:

N = Var \*100 , μ = Var , 10 σ2 = Var , де Var – номер варіанта, N – кількість дослідів, μ – математичне сподівання, σ2 – дисперсія.

4 Зберегти отриману вибірку у форматі .xls (Exсel).

5 Превірити гіпотезу про нормальний розподіл вибірки, використовуючи критерій Пірсона – χ2 і критерій Колмогорова з використанням внутрішніх функцій мови R.

6 Зробити висновок.

7 Оформити звіт.

8 Відповісти на контрольні запитання.

**Хід виконання лабораторної роботи:**

Вхідні дані:

* Var = 15 - номер варіанта;
* N = 15 \* 100 - кількість дослідів;
* µ = Var = 15 - математичне очікування;
* σ2 = Var/10 = 1,5- дисперсія.

**Аналіз отриманих результатів у Statgraphics**

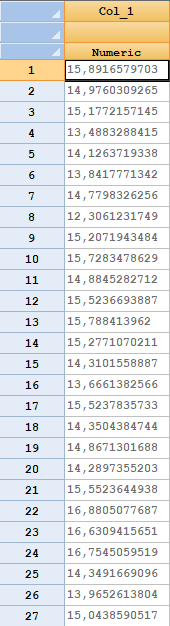


Рисунок 2.1 – Вибірка отримана за допомогою пакета Statgraphics

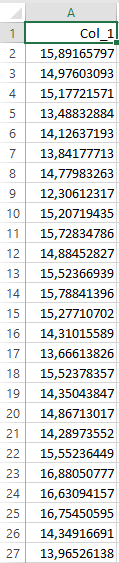


Рисунок 2.2 – Збережена вибірка у форматі .xls

Для аналізу результатів вибірки у пакеті Statgraphics було використано Distribution Fitting, що дозволяє побачити аналіз вибірки, гістограми(рис. 2.3-2.4), результати за тестом Шапіро-Уілка, тестом Колмогорова-Смірнова.

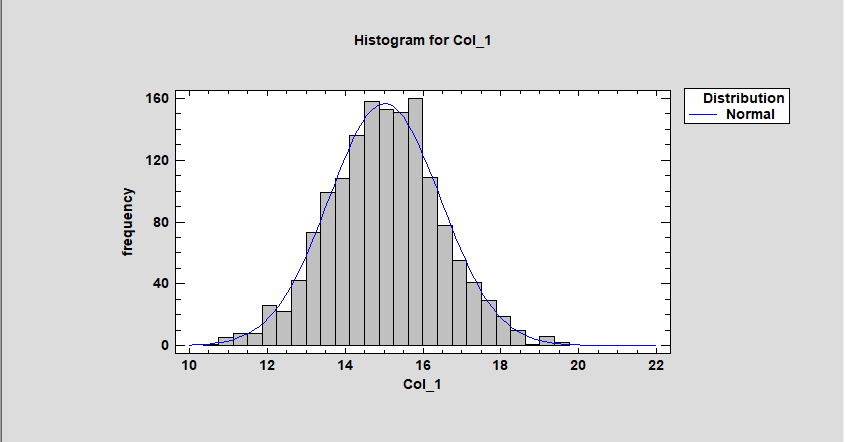


Рисунок 2.3 – Гістограма отриманих результатів

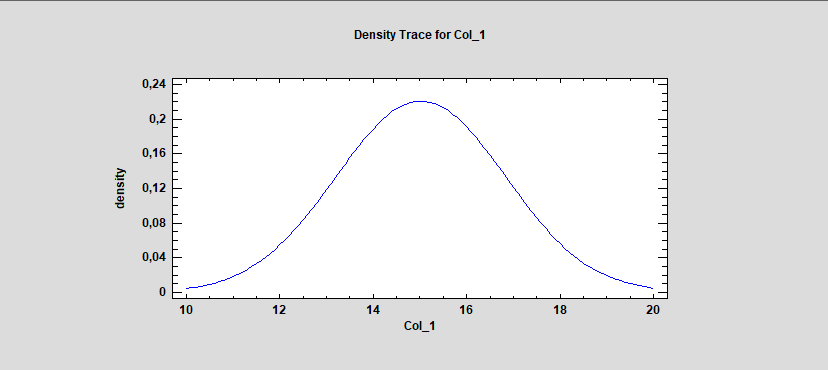


Рисунок 2.4 – Гістограма за щільністю

Тест Шапіро-Уілка використовується для перевірки гіпотези про нормальність розподілу вибірки. Його основна мета — визначити, чи відповідає даний набір даних нормальному розподілу. Якщо значення p-value більше рівня значущості (0,05), то вибірка розглядається як така, що відповідає нормальному розподілу.

З результатів тесту для нашої вибірки (рис. 2.5) можна побачити наступне: p-value = 0,573554, що набагато більше за 0,05. Тому ми не відкидаємо нульову гіпотезу, яка наголошує на нормальності розподілу. З 95% впевненістю можна сказати, що дані вибірки підпорядковуються нормальному розподілу.

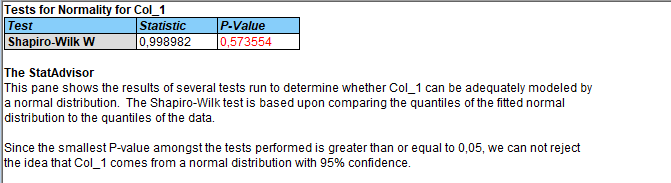


Рисунок 2.5 – Тест Шапіро-Уілка

Тест Колмогорова-Смірнова показує наскільки емпірична функція розподілу відхиляється від теоретичної (в даному випадку нормальної) функції розподілу.

З результатів тесту для нашої вибірки (рис. 2.6) можна побачити наступне: DPLUS – максимальне позитивне відхилення емпіричної функції розподілу від теоретичної, що дорівнює 0.0157472. DMINUS – максимальне негативне відхилення емпіричної функції розподілу від теоретичної, що дорівнює 0.0136922. DN – загальне максимальне відхилення між емпіричною функцією розподілу і теоретичною функцією нормального розподілу, тобто найбільше зі значень DPLUS і DMINUS. Чим менші ці значення, тим менше відхилення від теоретичного нормального розподілу. В данному випадку всі значення досить малі, що свідчить про близькість емпіричної функції розподілу до нормальної. P-Value = 0.850926, що більше за 0,05. Тому ми не відкидаємо нульову гіпотезу, яка стверджує, що розподіл є нормальним.

Отже, 95% впевненістю ми можемо стверджувати, що розподіл вибірки відповідає нормальному розподілу.

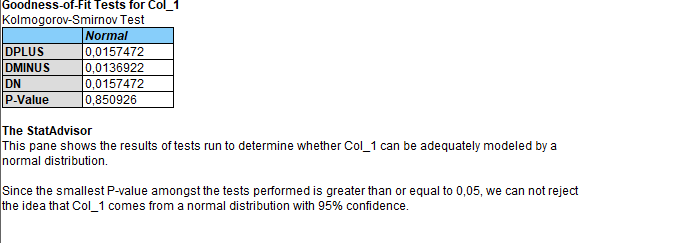
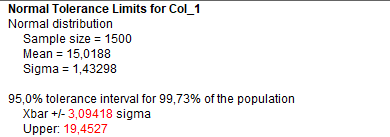


Рисунок 2.6 – Тест Колмогорова-Смірнова

Normal Tolerance Limits (Нормальні толерантні межі) передбачають, що дані слідують нормальному розподілу. Тому використання цих меж саме по собі припускає нормальність розподілу



arr <- array(rnorm(1500, 15,1.5),dim = c(1500,1),dimnames = list(c(1:1500)))

hist(arr, main = "Histogram", col='purple')

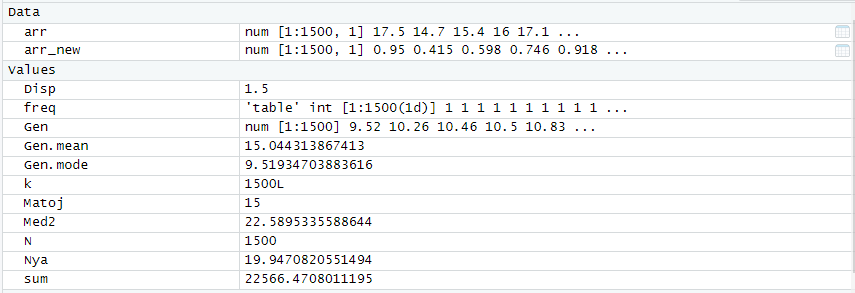
plot(density(arr), main = "Plot")

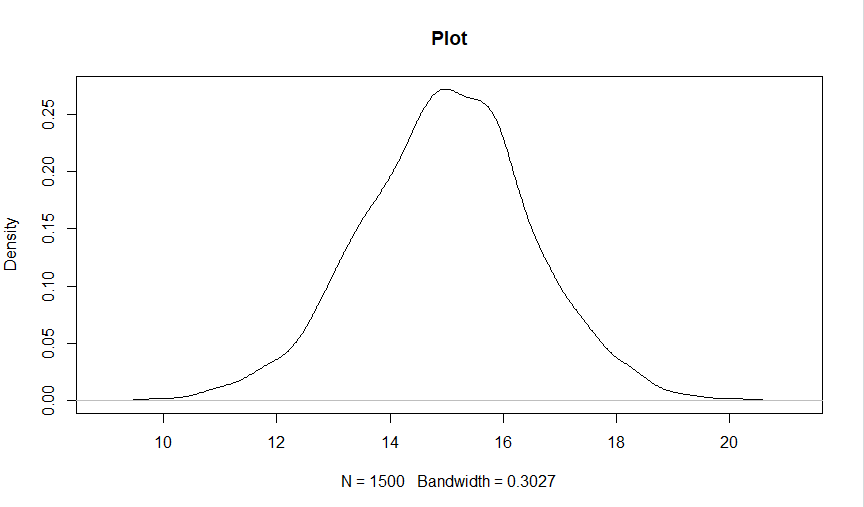
shapiro.test(arr) #Шапіро-Уілка тест

arr\_new <- pnorm(arr, 15, 1.5) #Масив ймовірностей розподілення

chisq.test(arr\_new) #Критерій узгодження Пірсона

ks.test(arr, arr\_new) # Тест Колмагорова-Смірнова





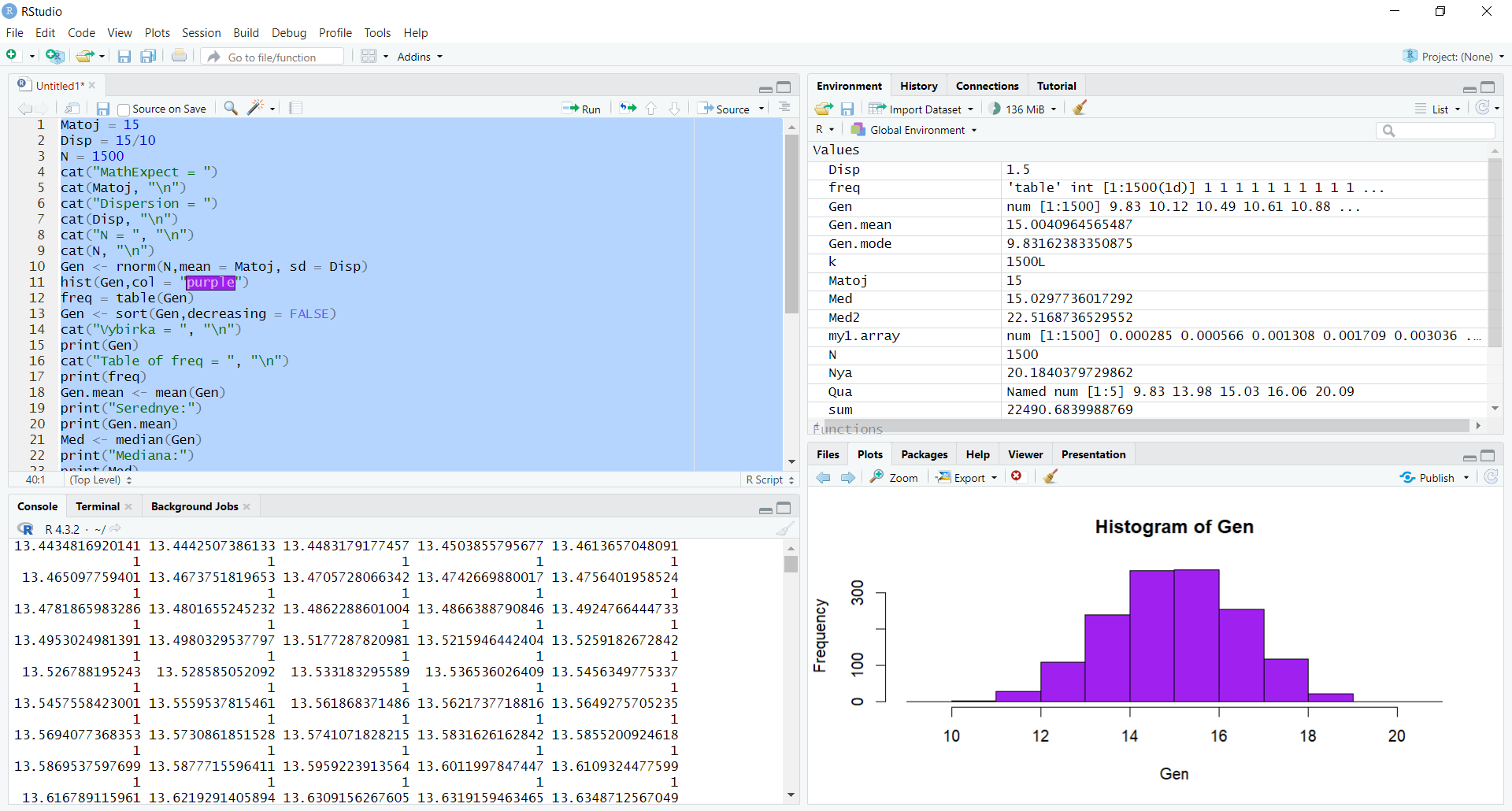


Рисунок 2.2 – робота з RStudio

Побудуємо гістограму ряду розподілу, отримаємо середнє арифметичне, медіану, моду, розмах (у відсотках), таблицю частот, ряд розподілу, наведений код нижче (рисунок 2.3) містить необхідні рядки для компіляції.

**Код програми, використаний під час роботи у RStudio:**

Matoj = 15

Disp = 15/10

N = 1500

cat("MathExpect = ")

cat(Matoj, "\n")

cat("Dispersion = ")

cat(Disp, "\n")

cat("N = ", "\n")

cat(N, "\n")

Gen <- rnorm(N,mean = Matoj, sd = Disp)

hist(Gen,col = "purple")

freq = table(Gen)

Gen <- sort(Gen,decreasing = FALSE)

cat("Vybirka = ", "\n")

print(Gen)

cat("Table of freq = ", "\n")

print(freq)

Gen.mean <- mean(Gen)

print("Serednye:")

print(Gen.mean)

Med <- median(Gen)

print("Mediana:")

print(Med)

Mode <-function(x){

ux <- unique(x)

ux[which.max(tabulate(match(x, ux)))]

}

Gen.mode <- Mode(Gen)

print("Mode:")

print(Gen.mode)

Qua <- quantile(Gen)

print("Quantile:")

print(Qua)

hist(Gen,main = "Hiostogram of observed data")

plot(density(Gen), main = "Density estimate of data")

shapiro.test(Gen)

my1.array <- pnorm(Gen, mean = Matoj, sd = Disp)

chisq.test(my1.array)

ks.test(Gen,my1.array)

Відкомпілюємо першу частину коду для побудови гістограми нормального розподілу (рис. 2.4).

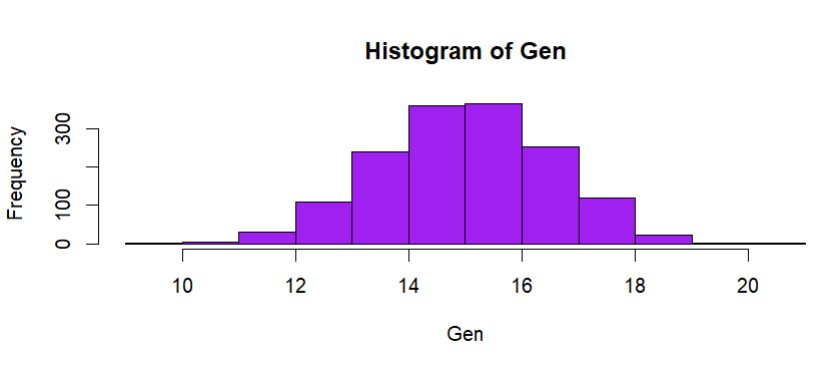


Рисунок 2.4 – гістограма нормального розподілу

Отримаємо необхідні значення у вікні із результатами компілювання коду (рис. 2.5)

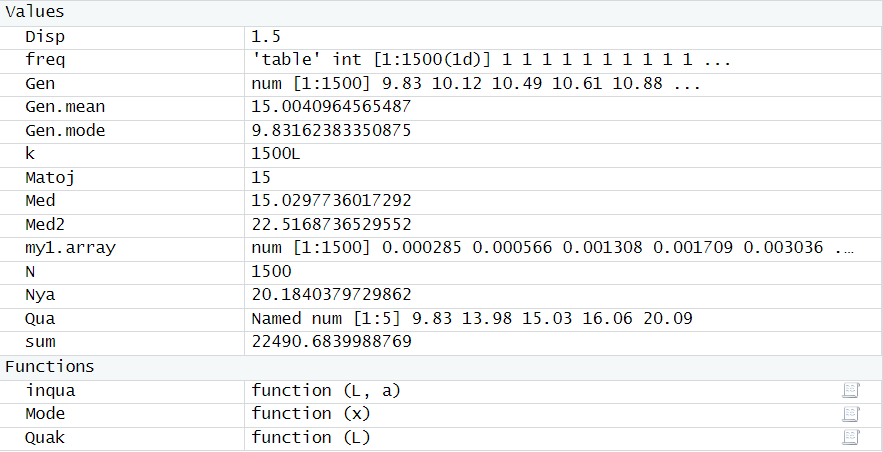


Рисунок 2.5 – результати побудови гістограми нормального розподілу

Під час роботи з кодом та внутрішніми функціями RStudio були виведені різні дані: вибірка (рис. 2.6), таблицю частот (рис. 2.7), середнє арифметичне, медіану, моду та квантиль (рис. 2.8) й відповідні дані як результат роботи програми (рис. 2.9).

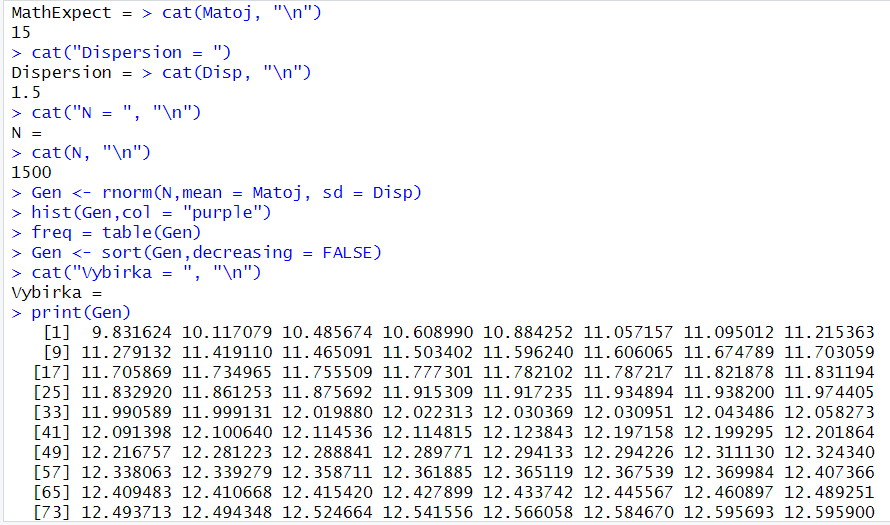
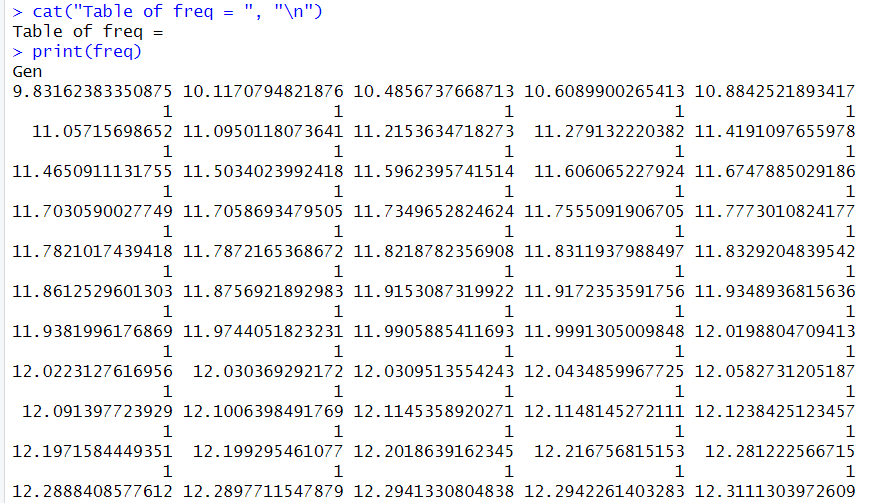


Рисунок 2.6 – отримана вибірка за допомогою мови програмування R

  
Рисунок 2.7 – отримана таблиця частот

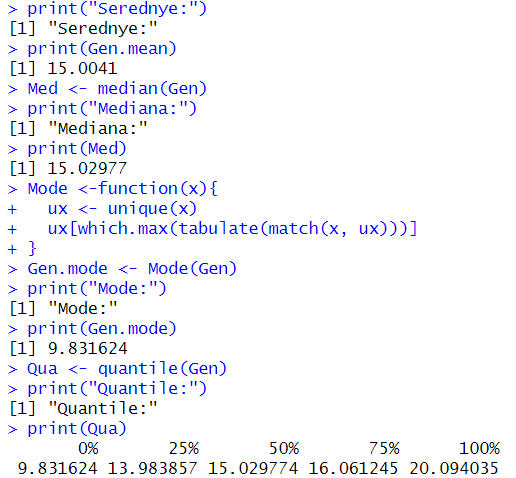


Рисунок 2.8 – отримані середнє арифметичне, медіана, мода та квантиль

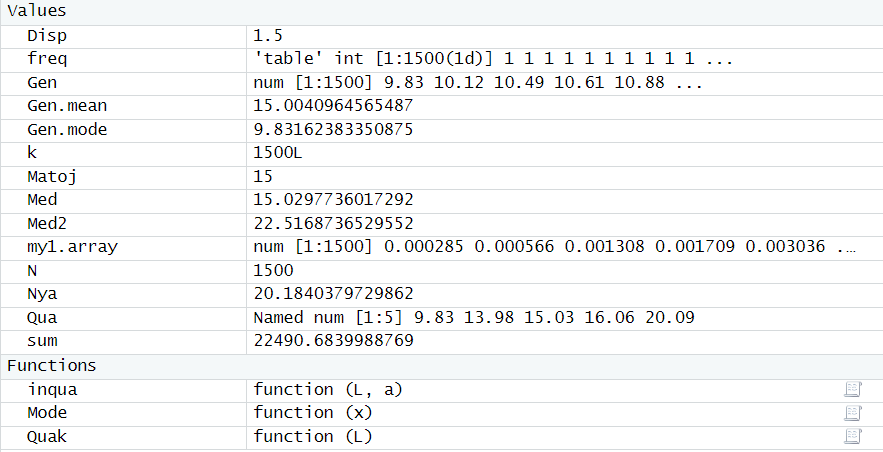


Рисунок 2.9 – отримані характеристики результату компіляції коду

Для перевірки гіпотези про нормальний розподіл вибірки, використовуючи критерій Пірсона, Колмогорова-Смірнова та Шапіро-Уілка треба записати відповідний код й відкомпілювати його, це є ключовою умовою побудови гістограми за щільністю (рис. 2.10) та виведення необхідного результату після проведення аналізу вибірки (рис. 2.11).

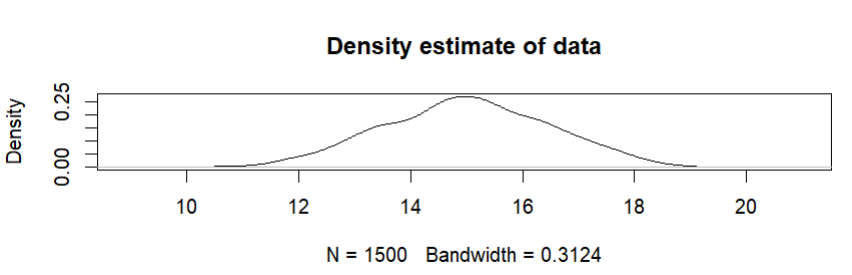


Рисунок 2.10 – гістограма за щільністю

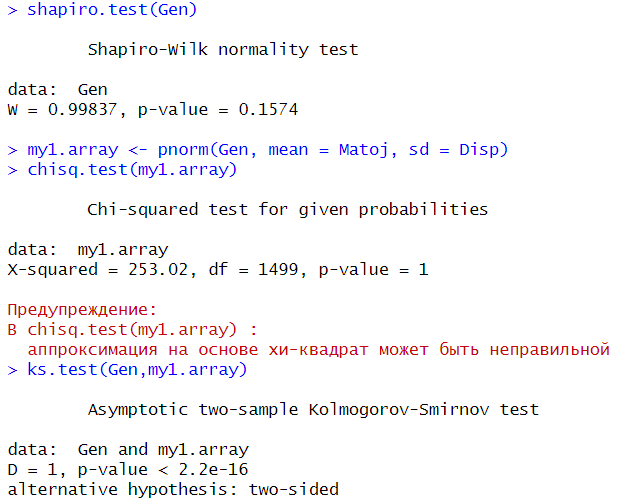


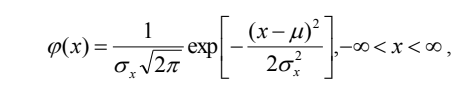
Рисунок 2.11 – результати перевірки гіпотез

**Контрольні запитання**

1. Які ви знаєте параметри нормального закону розподілу випадкових величин?

Математичне очікування µ та середнє квадратичне відхилення.

1. Наведіть аналітичний вираз щільності ймовірності для нормального закону розподілу.



1. Накресліть функцію щільності ймовірності для нормального закону.

Функція щільності ймовірності для нормального закону зображена на рисунку 2.12

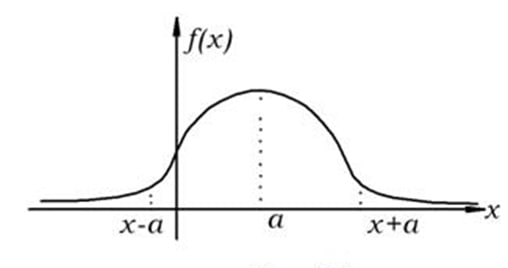


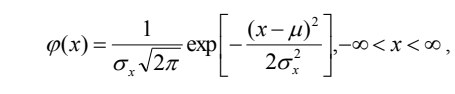
Рисунок 2.12 – функція щільності ймовірності для нормального закону

1. Чому дорівнюють коефіцієнт асиметрії та коефіцієнт ексцесу для нормального закону?

Коефіцієнт асиметрії та коефіцієнт ексцесу для нормального закону дорівнюють 0.

1. Яким критерієм визначається закон розподілу випадкової величини? Записати аналітичний вираз.
2. Для перевірки гіпотези про розподіл випадкової величини

використовують критерій Пірсона, критерій Колмагорова та ін.



1. Як змінюється вигляд гістограми при зміні величини інтервалу?

Якщо інтервал зменшити, то:

* стовпці гістограми стають вищими і більш вузькими;
* гістограма може більше відображати дрібні варіації в даних, і ми бачимо більше деталей;
* гістограма може виглядати більш "кусковою" і має більше стовпців.

Якщо інтервал збільшити, то:

* стовпці гістограми стають нижчими і ширшими;
* гістограма загальної форми може бути більш вирівняною і гладкою;
* ми втрачаємо деталі даних, і гістограма може дати загальну інформацію про розподіл без деталей.

1. Як використовується правило трьох сігм для визначення закону розподілу випадкової величини?

Закон трьох сигм полягає в тому, що, якщо випадкова величина розподілена нормально, то абсолютна величина її відхилення від математичного очікування не перебільшує потроєного середнього квадратичного відхилення.

На практиці правило трьох сигм застосовують так: якщо розподіл випадкової величини є невідомим, але умова, що вказана в наведеному прикладі виконується, то існує підстава вважати, що випадкова величина розподілена нормально; в іншому випадку вона не розподілена нормально.

1. Необхідність застосування статистичних методів обробки результатів спостережень.

Застосування статистичних методів обробки результатів спостережень має важливе значення у багатьох наукових, інженерних, медичних, соціальних та інших дисциплінах. Ось кілька ключових аспектів, які пояснюють необхідність статистичних методів:

* зведення до розуміння: статистичні методи допомагають аналізувати та розуміти дані;
* прийняття рішень: в багатьох випадках рішення, які приймаються на основі даних, повинні бути обґрунтовані статистичними доказами;
* попередження помилкових висновків: статистичні методи допомагають визначити, наскільки значущими є отримані результати;
* перевірка гіпотез: статистичні методи дозволяють науковцям та дослідникам перевіряти гіпотези, які стосуються взаємозв'язків між змінними;
* прогнозування та моделювання: статистичні методи використовуються для розробки прогнозних моделей та аналітики, яка допомагає передбачити майбутні події на основі історичних даних;
* управління ризиками: статистика допомагає визначити ризики та вирішити, як їх управляти;
* дослідження соціальних явищ: статистичні методи дозволяють аналізувати соціальні, економічні і політичні явища.

1. У чому полягає основний принцип сатистичних гіпотез?

Основний принцип статистичних гіпотез полягає в тому, що дослідники ставлять дві зворотні гіпотези про певний параметр або явище, і потім використовують статистичні методи для визначення, яка з цих гіпотез більш вірогідна або підтверджена даними.

1. Які гіпотези можна перевірити за допомогою критерія Пірсона, Колмогорова, Фішера, Стьюдента, Кохрена?

*Критерій Пірсона (χ²):*

Перевірка гіпотези про асоціацію (незалежність) між категоріальними змінними в таблицях спряженості (критерій хі-квадрат).

Перевірка гіпотези про рівність розподілу категоріальної змінної спостережень та очікуваних частот.

*Критерій Колмогорова-Смірнова:*

Перевірка гіпотези про відповідність емпіричного розподілу даних теоретичному розподілу (нормальному або іншому).

*Критерій Фішера:*

Перевірка гіпотези про рівність дисперсій між двома або більше групами в аналізі дисперсії (ANOVA).

Перевірка гіпотези про ефективність нового методу лікування або впливу фактору на якість результатів.

*Критерій Стьюдента:*

Перевірка гіпотези про різницю між середніми двох груп (незалежні вибірки) за допомогою t-критерію Стьюдента.

Перевірка гіпотези про різницю між середніми у випадку залежних (спарених) вибірок.

*Критерій Кохрена:*

Перевірка гіпотези про рівність дисперсій у випадку аналізу дисперсії перед проведенням подальшого аналізу.

1. Як визначити довірчі інтервали для математичного сподівання?

Довірчий інтервал для математичного сподівання (середнього значення) визначається за допомогою статистичних методів на основі вибірки даних.

Кроки для визначення довірчого інтервалу для математичного сподівання:

* Зібрати вибірку даних;
* Обчислити середнє значення та стандартну помилку;
* Вибрати рівень довіри;
* Визначити критичні значення;
* Визначити довірчий інтервал;
* Інтерпретація результатів.

Висновки: під час виконання лабораторної роботи на основі розробленого коду були отримані основні характеристики розподілу ймовірностей випадкової величини та побудована гістограма за щільністю, яка допомогла перевірити гіпотезу про закон розподілу вибірки за допомогою різних критеріїв, такі як: Пірсона, Колмогорова-Смірнова та Шапіро-Уілка. Цей аналіз допоміг ознайомитися з можливостями пакетів статистичної обробки даних, закріпивши свої знання математичної статистики, використовуючи вбудовані функції мови програмування R.