Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**Отчёт лабораторная №6**

**Дисциплина: Обработка больших данных**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Гаранина Л.В.

Направление подготовки 02.03.02 Фундаментальная информатика и

информационные технологии

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Яхонтов А.А.

Краснодар

2025

**Тема:** Проверка статистических гипотез.

**Цель**: ознакомиться с некоторыми статистическими тестами, принципами их работы. Научиться оценивать нормальность распределения выборки, а также выполнить оценку статистических гипотез.

**Задание:**

1. Провести дескриптивный анализ данных.

2. Провести проверку на нормальность и дисперсию. Из чего сделать вывод о требуемом тесте.

3. Проверьте гипотезу о среднем весе спортсменов выбранного вида спорта (вид спорта остается по ЛР4).

4. Проверьте гипотезу о равенстве среднего веса женщин (мужчин) в двух разных выбранных видах спорта (сравнение двух независимых выборок – двухвыборочный критерий.).

Вид спорта: фигурное катание.

**Ход работы**:

Создадим датасет, в который загрузим данные о фигурном катании из файла с набором данных о современных Олимпийских играх. Датасет показан на рисунке 1.



Рисунок 1 – Датасет о фигурном катании

Проведем дескриптивный анализ данных. Для этого создадим датасет с возрастом, ростом и весом. В дескриптивный анализ входит мода, медиана и среднее. Результаты анализа показаны на рисунке 2.

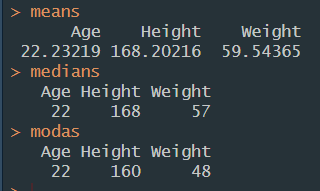


Рисунок 2 – Мода, медиана и среднее

Также построим боксплоты и гистограммы, показанные на рисунке 3.

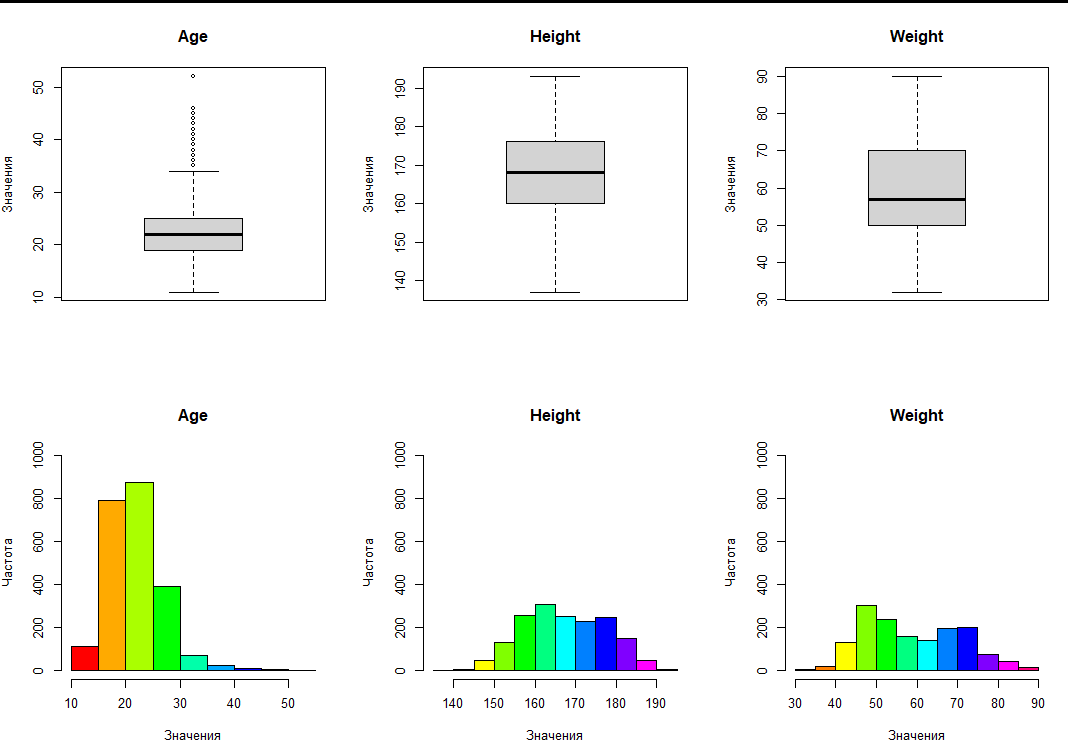


Рисунок 3 – Боксплоты и гистограммы

Теперь проведем проверку на нормальность и дисперсию. Это необходимо, чтобы выбрать корректный тест, который даст достоверные результаты.

Построим гистограмму с нормальной кривой плотности и квантильно-квантильные графики (базовый и из пакета car). Первый показывает насколько данные соответствуют нормальному распределению, второй – распределение данных относительно ожидаемого нормального распределения. Все графики показаны на рисунке 4.

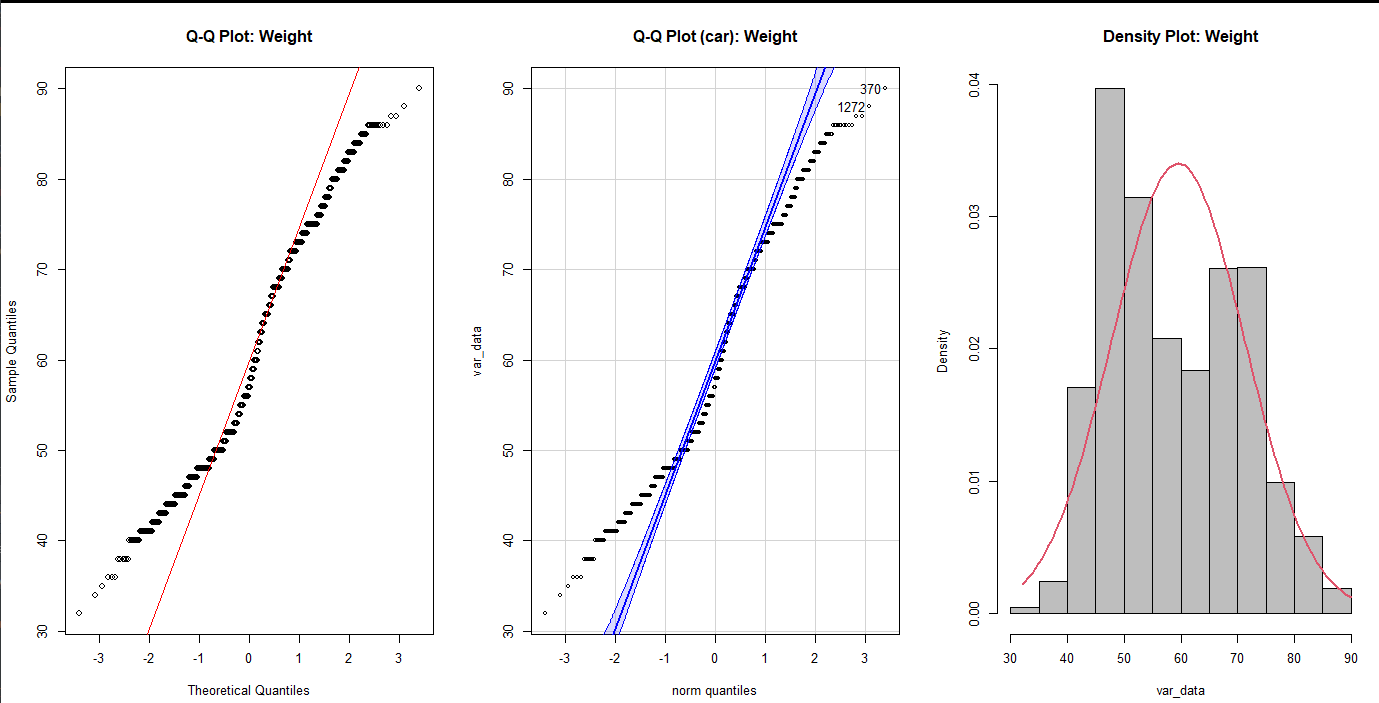


Рисунок 4 – Графики проверка на нормальность

Q-Q график сравнивает квантили реального распределения с квантилями нормального распределения. Видим, что точки в центре лежат вдоль линии, но на концах наблюдаются отклонения. На 2 графике то же самое – много точек выходят за пределы доверительных интервалов – значительное отклонение от нормальности. На 3 графике показана гистограмма распределения веса, где красная линия – теоретическая кривая нормального распределения, построенного по среднему и стандартному отклонению данных. Можем заметить, что гистограмма асимметрична и имеет перекос вправо. Все графики показывают, чт распределение веса фигуристов не является нормальным.

Также проедем тест Шапиро, чтобы еще раз убедиться в своих выводах. Это показано на рисунке 5.

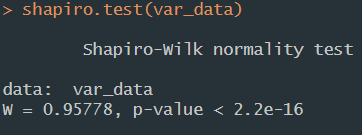


Рисунок 5 – Результаты теста Шапиро

По тесту Шапиро видно, что p-value < 0.05 – статистически значимое отклонение от нормального распределения. Это значит, что распределение веса фигуристов не является нормальным, то есть мы отвергаем нулевую гипотезу о нормальности и далее можно применять только непараметрические тесты, которые не чувствительны к выбросам и асимметрии.

Проверим также равенство дисперсий между группами полов.

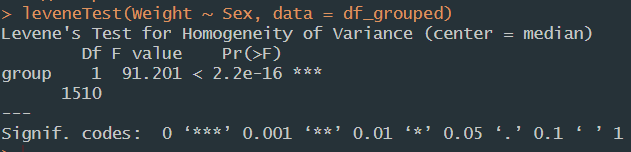


Рисунок 6 – Проверка дисперсии

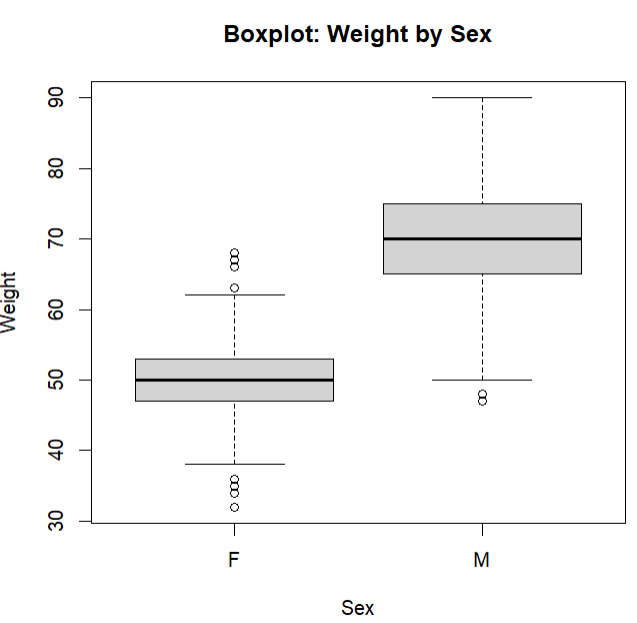


Рисунок 7 – Боксплоты дисперсий

Можем заметить, что дисперсии веса у мужчин и женщин статистически значимо отличается.

Предположим, что средний вес фигуристов равняется 65 – нулевая гипотеза. Проверим тестом Вилкоксона.

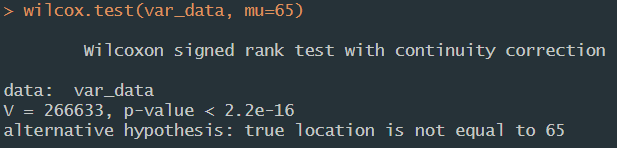


Рисунок 8 – Проверка 1 гипотезы

Можем заметить, что p-value < 0.05, а значит наша нулевая гипотеза является неверной. Далее проверим гипотезу, что средний вес фигуристов равен 59.

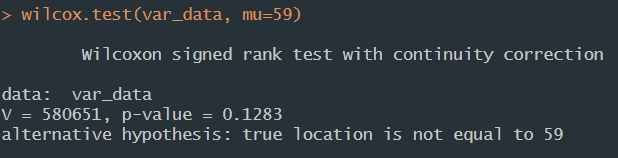


Рисунок 9 – Проверка 2 гипотезы

Значение p-value достаточно велико, а значит, что нулевая гипотеза о среднем весе спортсменов оказалась верной.

Теперь сравним вес женщин(мужчин) в фигурном катании и гимнастике. Для каждых проведем тест Шапиро. Результаты показаны на рисунке 10. Также построим графики 11-14.

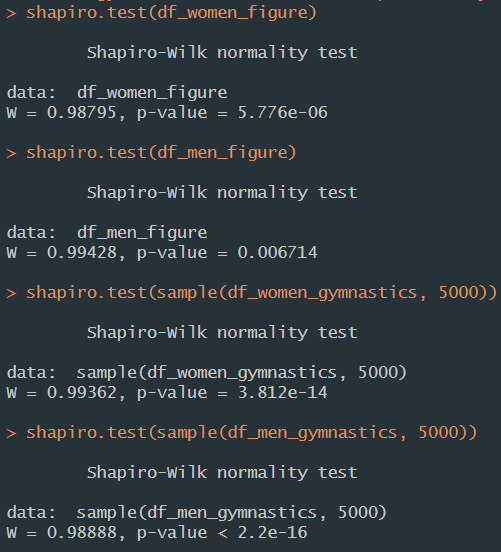


Рисунок 10 – Результаты теста Шапиро

По результатам теста ни у кого нет нормального распределения. Это же видно и по графикам ниже.

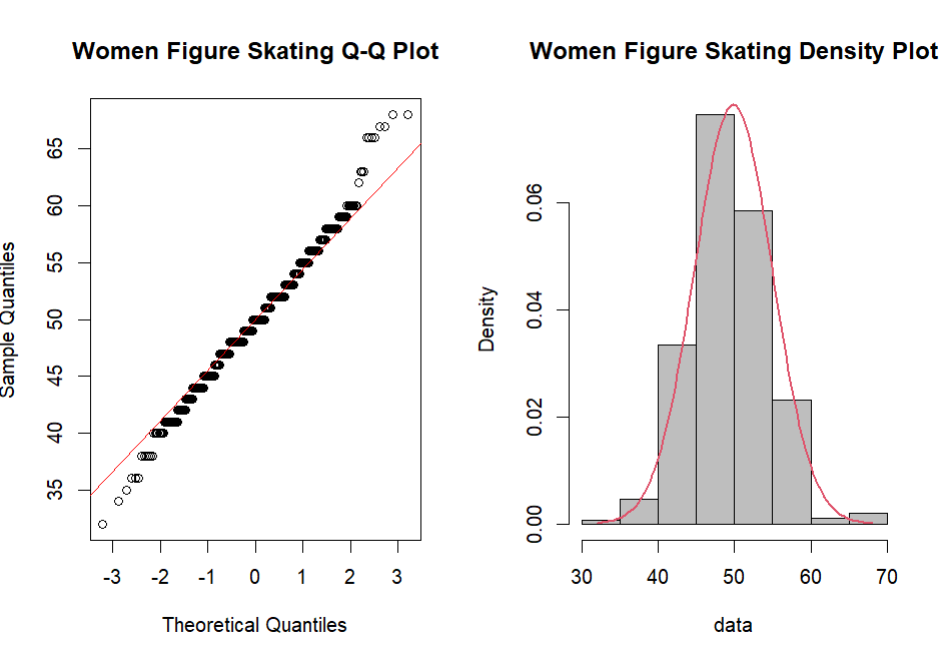


Рисунок 11 - Графики распределения женщин-фигуристок

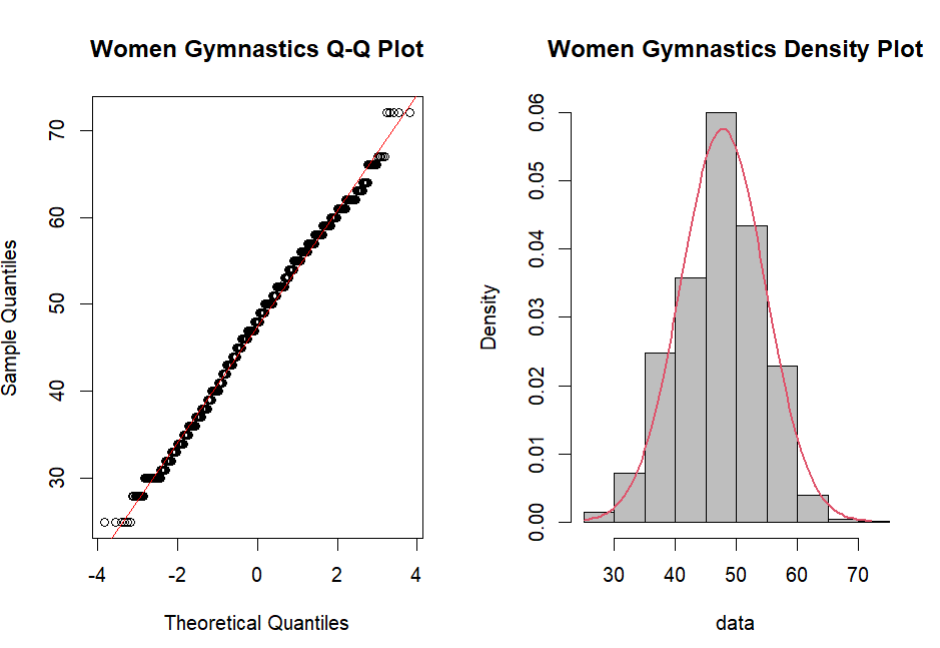


Рисунок 12 – Графики распределения женщин-гимнасток

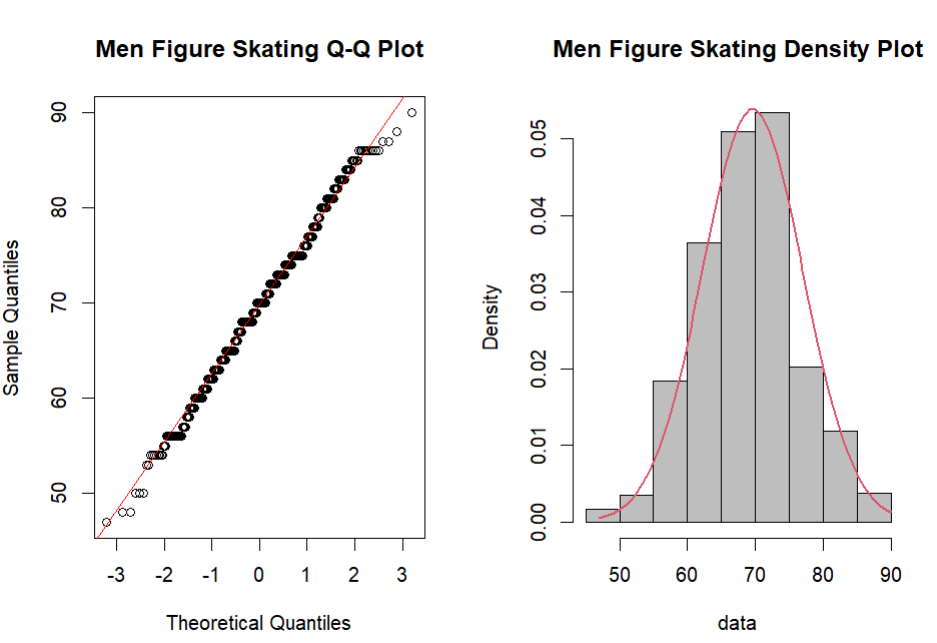


Рисунок 13 - Графики распределения мужчин-фигуристов

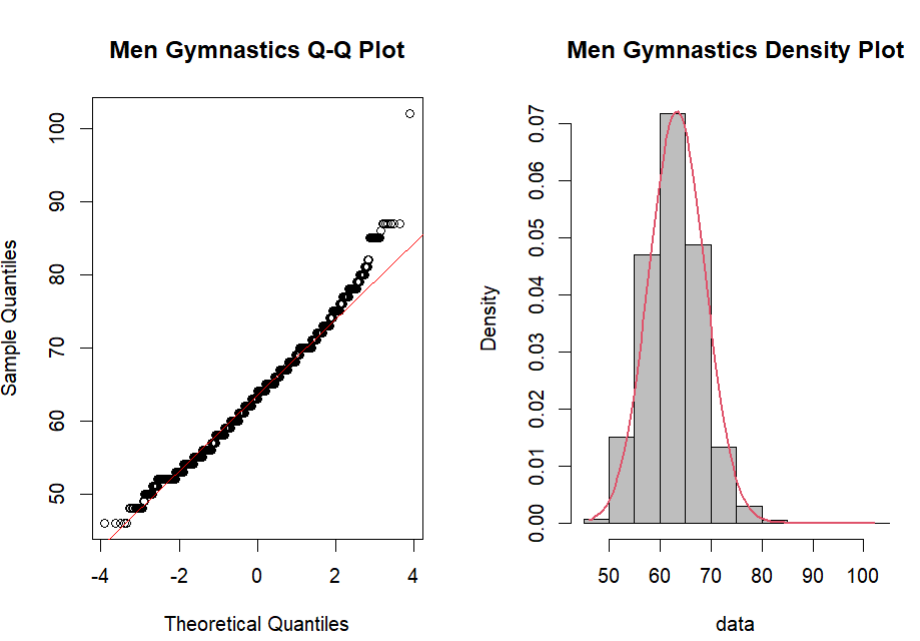


Рисунок 14 - Графики распределения мужчин-гимнастов

Так как распределение не является нормальным, то применим тесты Вилкоксона соотношения веса и спорта для женщин и мужчин. Этот тест использует сравнение медина 2-ух независимых выборок. Нулевая гипотеза – нет различий между группами по распределению веса по видам спорта.

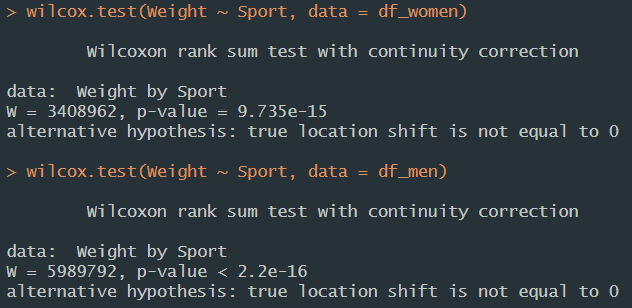


Рисунок 15 – Результаты теста Вилкоксона

Везде очень маленькое значение p-уровня, значит, мы отвергаем нулевую гипотезу и принимаем альтернативную – существует статистически значимое различие в весе между мужчинами и женщинами в разных видах спорта.

**Вывод:** проделав лабораторную работу, были изучены некоторые статистические тесты, а также оценена нормальность выборок и оценены статистические гипотезы.

**Листинг программы:**

library(car)

setwd("C:/Users/Людмила/Desktop/big data/rstudioLAB")

df<-read.csv("athlete\_events.csv", sep=",", header=TRUE)

df\_figure <- df[df$Sport == "Figure Skating", ]

# 1. Дескриптивный анализ

df\_numeric <- df\_figure[, c("Age", "Height", "Weight")]

summary(df\_numeric)

means = apply(df\_numeric, 2, function(x) mean(x, na.rm = TRUE))

medians = apply(df\_numeric, 2, function(x) median(x, na.rm = TRUE))

getmode <- function(v) {

v <- na.omit(v)

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

modas = apply(df\_numeric, 2, getmode)

par(mfrow = c(2, 3))

for (i in 1:ncol(df\_numeric)) {

boxplot(df\_numeric[, i],

main = colnames(df\_numeric)[i],

ylab = "Значения")

}

for (i in 1:ncol(df\_numeric)) {

h <- hist(as.numeric(df\_numeric[[i]]), plot = FALSE)

bar\_colors <- rainbow(length(h$counts))

plot(h,

main = colnames(df\_numeric)[i],

xlab = "Значения",

ylab = "Частота",

col = bar\_colors,

ylim = c(0, 1000))

}

# 2. Провести проверку на нормальность и дисперсию. Из чего сделать вывод о требуемом тесте.

# Q-Q графики

df\_grouped <- df\_figure[, c("Sex", "Age", "Weight")]

df\_grouped$Sex <- as.factor(df\_grouped$Sex)

df\_grouped <- na.omit(df\_grouped)

par(mfrow = c(1, 3))

var\_data <- df\_grouped$Weight

# Q-Q график (базовый)

qqnorm(var\_data, main = "Q-Q Plot: Weight")

qqline(var\_data, col = "red")

# Q-Q график (из пакета car)

qqPlot(var\_data, main = "Q-Q Plot (car): Weight")

# Гистограмма с нормальной кривой плотности

x2 <- seq(min(var\_data), max(var\_data), length.out = 100)

fun <- dnorm(x2, mean = mean(var\_data), sd = sd(var\_data))

hist(var\_data, freq = FALSE, col = "gray", main = "Density Plot: Weight")

lines(x2, fun, col = 2, lwd = 2)

set.seed(0)

shapiro.test(var\_data)

# Дисперсия

leveneTest(Weight ~ Sex, data = df\_grouped)

boxplot(Weight ~ Sex, data = df\_grouped, main = "Boxplot: Weight by Sex", ylab = "Weight")

# 3.

wilcox.test(var\_data, mu=65)

wilcox.test(var\_data, mu=59)

# 4.

df\_selected\_sports <- df[df$Sport %in% c("Figure Skating", "Gymnastics"), c("Sex", "Age", "Height", "Weight", "Sport")]

df\_selected\_sports <- na.omit(df\_selected\_sports)

df\_women <- df\_selected\_sports[df\_selected\_sports$Sex == "F", ]

df\_men <- df\_selected\_sports[df\_selected\_sports$Sex == "M", ]

df\_women\_figure = df\_women[df\_women$Sport == "Figure Skating",]$Weight

df\_women\_gymnastics = df\_women[df\_women$Sport == "Gymnastics",]$Weight

df\_men\_figure = df\_men[df\_men$Sport == "Figure Skating",]$Weight

df\_men\_gymnastics = df\_men[df\_men$Sport == "Gymnastics",]$Weight

shapiro.test(df\_women\_figure)

shapiro.test(df\_men\_figure)

shapiro.test(sample(df\_women\_gymnastics, 5000))

shapiro.test(sample(df\_men\_gymnastics, 5000))

plot\_graphs <- function(data, title\_prefix) {

par(mfrow = c(1, 2))

# Q-Q график

qqnorm(data, main = paste(title\_prefix, "Q-Q Plot"))

qqline(data, col = "red")

# Гистограмма с нормальной кривой

x2 <- seq(min(data), max(data), length.out = 100)

fun <- dnorm(x2, mean = mean(data), sd = sd(data))

hist(data, freq = FALSE, col = "gray", main = paste(title\_prefix, "Density Plot"))

lines(x2, fun, col = 2, lwd = 2)

}

plot\_graphs(df\_women\_figure, "Women Figure Skating")

plot\_graphs(df\_women\_gymnastics, "Women Gymnastics")

plot\_graphs(df\_men\_figure, "Men Figure Skating")

plot\_graphs(df\_men\_gymnastics, "Men Gymnastics")

wilcox.test(Weight ~ Sport, data = df\_women)

wilcox.test(Weight ~ Sport, data = df\_men)