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**

**ЛАБОРАТОРНАЯ РАБОТА №7**

**Дисциплина: Обработка больших данных**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Придава А.А.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Шиян В.И.

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

Из исходного csv-файла были импортированы данные в RStudio. Таблица представлена на рисунке 1.

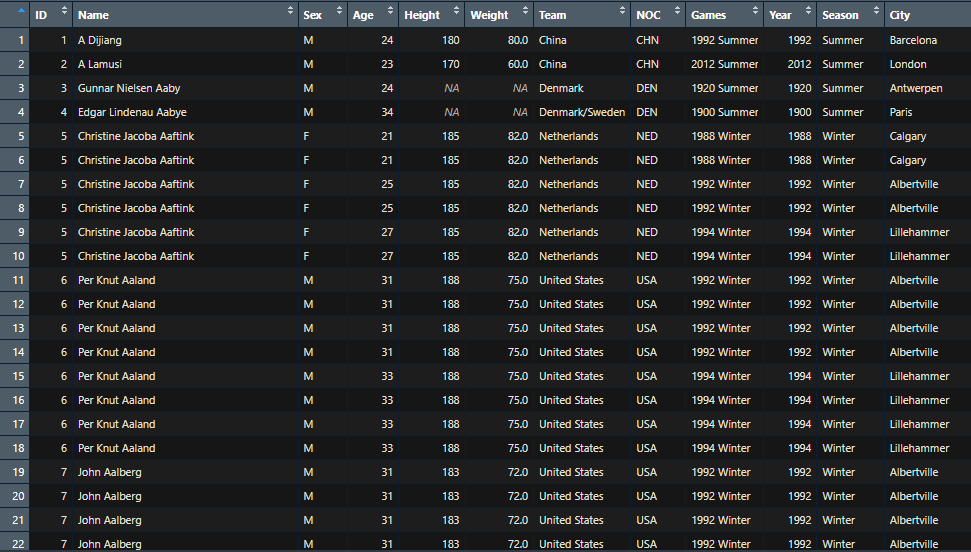


Рисунок 1 – Исходная таблица с данными

В данном датасете представлены данные обо всех спортсменах на всех Олимпийских играх с 1896 года по 2016 год. Удалим всех спортсменов, кроме спортсменов выбранного вида спорта, а также повторяющиеся строчки, и строчки с пустым значением поля “Вес”. На рисунке 2 изображена таблица получившихся данных.



Рисунок 2 – Измененная таблица с данными

Затем проведем тест Шапиро-Уилкса на нормальность распределения данных в полученной таблице. В качестве нулевой гипотезы берется утверждение о том, что данные распределены нормально. На рисунке 3 изображен результат работы теста.

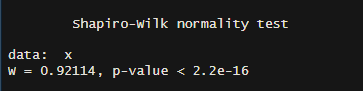


Рисунок 3 – Результат теста Шапиро-Уилкса

Значение p-value < 2.2e-16. Так как это значение меньше, чем граничное значение 0.05, выборка данных не имеет нормального распределения на уровне значимости 0.05 (p-value < 0.05). Визуальное подтверждение изображено на рисунках 4-5.

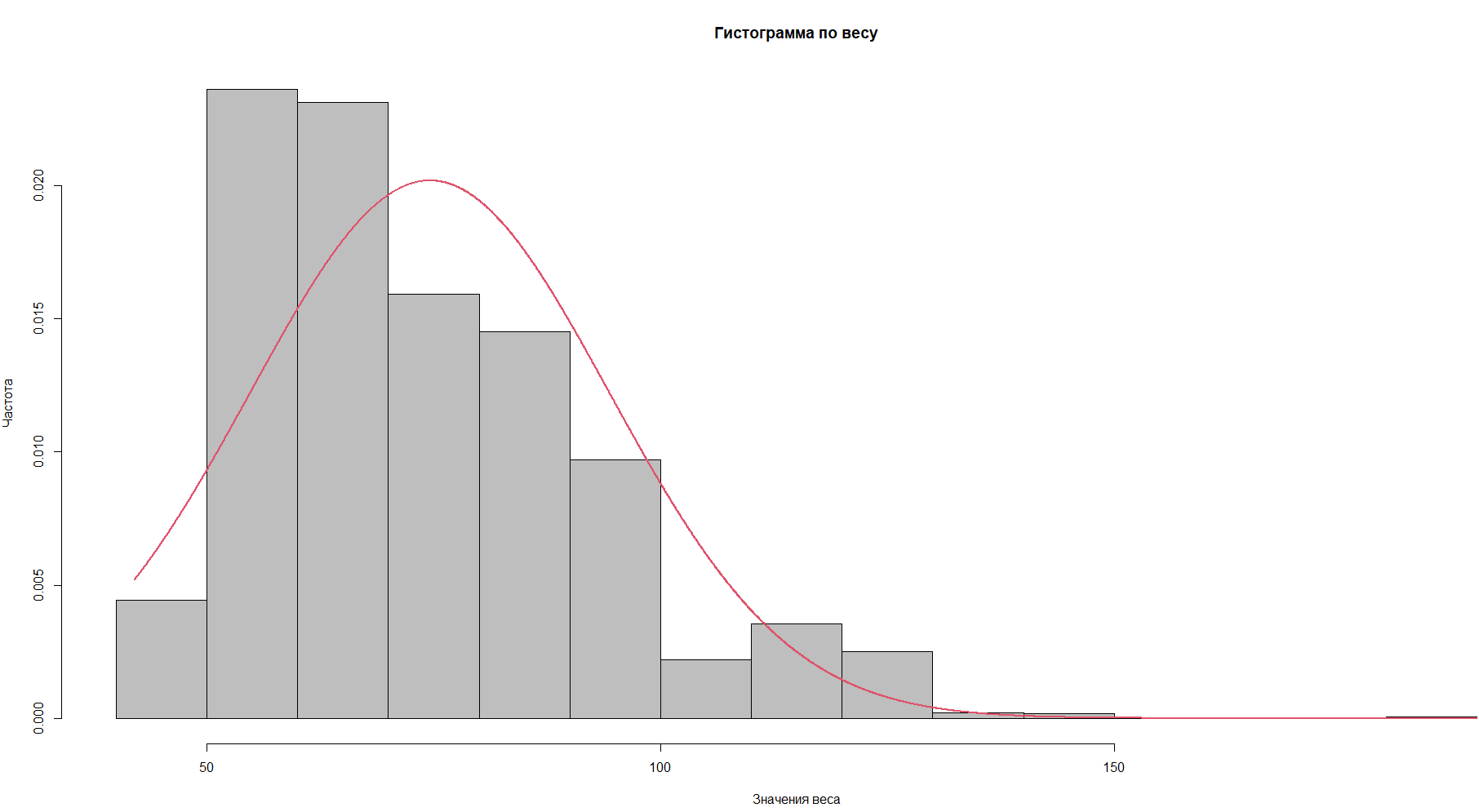


Рисунок 4 – Гистограмма с линией плотности

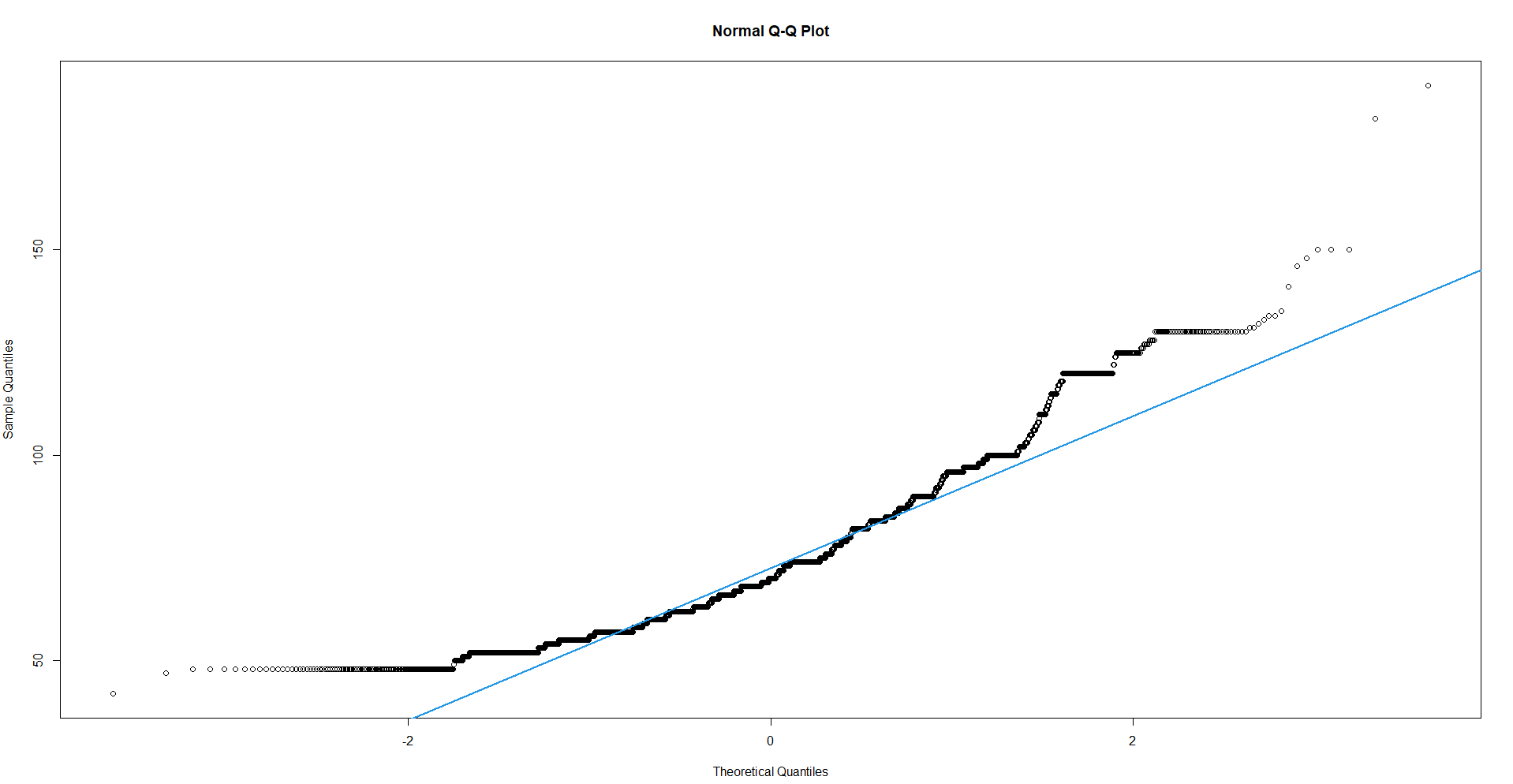


Рисунок 5 – Квантильно-квантильный график

Далее проверим гипотезу о среднем весе спортсменов, занимающихся борьбой. Первая гипотеза гласит: “Средний вес спортсменов, занимающихся борьбой, составляет 70 килограмм”. На рисунке 6 показан результат выполнения теста.

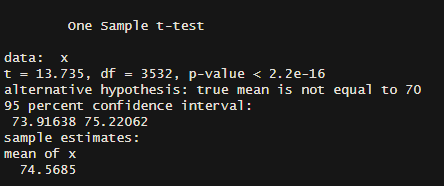


Рисунок 6 – Результат выполнения теста Стьюдента

Значение p-value < 0.05, поэтому гипотеза отвергается. Доверительный интервал расположен от 73.91 до 75.22 именно в этот интервал входит настоящий средний вес спортсменов – 74.56.

В следующем задании необходимо было проверить гипотезу о равенстве среднего веса спортсменок, занимающихся борьбой и гимнастикой.

Сперва вновь данные проверяются на нормальность распределения тестом Шапиро-Уилкса. На рисунке 7 представлен результат этого теста.

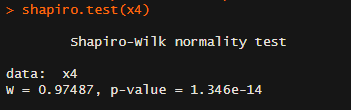


Рисунок 7 – Результат выполнения теста Шапиро-Уилкса

По результатам теста, значение p-value вновь меньше граничного 0.05. Следовательно, гипотеза о нормальном распределении не подтвердилась. Для большей наглядности построим графики для данных по борьбе и гимнастике. График изображён на рисунке 8.

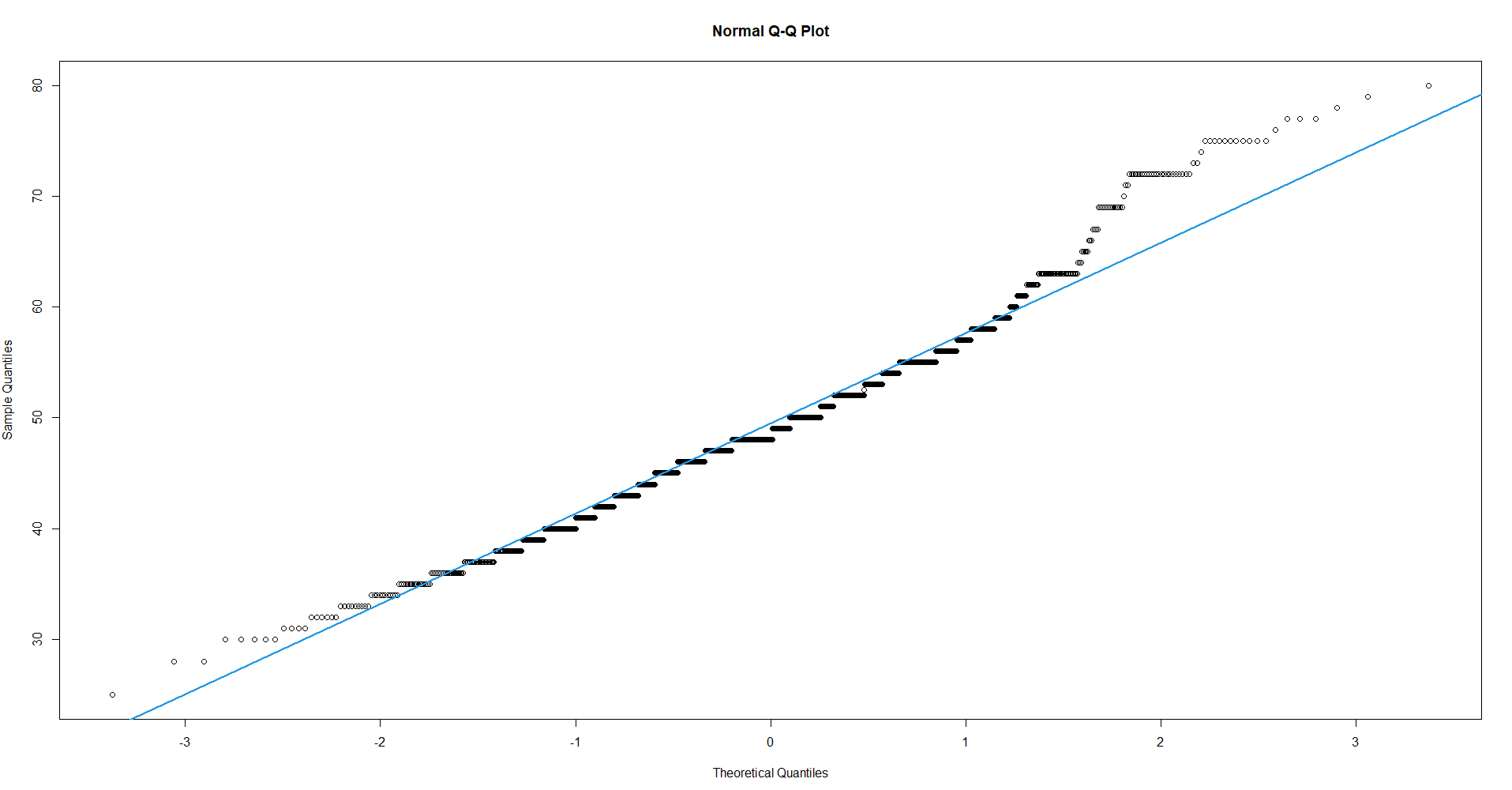


Рисунок 8 – Квантильно-квантильный график для двух видов спорта

График подтверждает, что данные не распределены нормально.

Затем, непосредственно проведем применим тест Флингера-Киллина для проверки равенства веса. Он непараметрический и отлично подходит для анализа данных с ненормальным распределением. На рисунке 9 изображён результат теста.

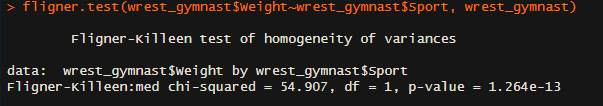


Рисунок 9 – Результаты выполнения теста Флингера-Киллина

Видно, что значение p-value близко к нулю, поэтому отвергаем нулевую гипотезу о равенстве веса. Для того, чтобы узнать приблизительные значения среднего веса спортсменок для двух видов спорта, применим двух выборочный тест Стьюдента. На рисунке 10 показан результат теста.

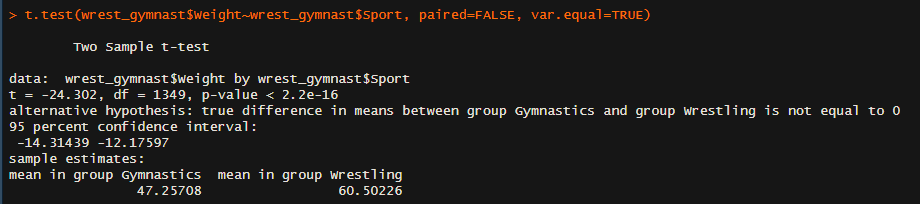


Рисунок 10 – Результат работы t-test

Значение p-value близко к нулю, а значит гипотеза не верна и равенство среднего веса спортсменок борьбы и гимнастики не подтверждено. Этот тест выводит также средний вес, поэтому можно наглядно убедиться, что средний вес спортсменок по гимнастике меньше на 13 килограммов веса спортсменок по борьбе.

**Вывод:** ознакомился с некоторыми статистическими тестами и принципами их работы, а также изучил оценивание нормальности распределения выборки и выполнение оценки статистических гипотез.

**Листинг программы**

# Импорт данных всех спортсменов по всем видам спорта

df <- read.table("C:/Users/vladc/OneDrive/Рабочий стол/Учёба/BigData/ЛР7/athlete\_events.csv", sep=",", header=TRUE)

wrestling<-df[which(df[, "Sport"]=="Wrestling"), c("Name", "Sex", "Weight", "Sport")]

# удаление строчек с пустым значением поля "Вес"

wrestling<-wrestling[-which(is.na(wrestling[, "Weight"])),]

# удаление повторяющихся строчек

wrestling<-unique(wrestling)

# значения столбца "Вес" становятся вектором

x<-wrestling[1:nrow(wrestling), "Weight"]

# Одномерные статистические тесты.

# проверка на нормальность распределения

# Тест Шапиро-Уилкса (Shapiro-Wilk test).

shapiro.test(x)

# Графический способ.

# гистограмма с линией плотности

x2<-seq(min(x), max(x), length=length(x))

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

hist(x, freq=FALSE, col="gray", main="Гистограмма по весу", xlab="Значения веса", ylab="Частота")

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

# квантильно-квантильный график

qqnorm(x)

qqline(x, col=4, lwd=2)

#title(main="Квантильно-квантильный график", xlab="Выборочные квантили", ylab="Теоретические квантили")

# Тест Стьюдента.

t.test(x, mu=70, conf.int=TRUE)

# Тест Уилкоксона.

wilcox.test(x, mu=mean(x), conf.int=TRUE)

#Задание 2.

wrest\_gymnast<-df[which(df[, "Sport"]%in%c("Gymnastics", "Wrestling")), c("Name", "Sex", "Weight", "Sport")]

wrest\_gymnast<-wrest\_gymnast[-which(is.na(wrest\_gymnast[, "Weight"])),]

wrest\_gymnast<-wrest\_gymnast[which(wrest\_gymnast[, "Sex"]=="F"),]

wrest\_gymnast$Sport <- factor(wrest\_gymnast$Sport)

wrest\_gymnast$Sport <- droplevels(wrest\_gymnast$Sport)

wrest\_gymnast <- unique(wrest\_gymnast)

x4<-wrest\_gymnast[1:nrow(wrest\_gymnast), "Weight"]

# проверка на нормальность распределения

# Тест Шапиро-Уилкса (Shapiro-Wilk test).

shapiro.test(x4)

# квантильно-квантильный график

qqnorm(x4)

qqline(x4, col=4, lwd=2)

# проверка равенство дисперсий

# Тест Флингера-Киллина.

fligner.test(wrest\_gymnast$Weight~wrest\_gymnast$Sport, wrest\_gymnast)

# проверка на отсутствие разницы в среднестатистическом значении

t.test(wrest\_gymnast$Weight~wrest\_gymnast$Sport, paired=FALSE, var.equal=TRUE)