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: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.М. Нагалевский

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В. И. Шиян

**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.

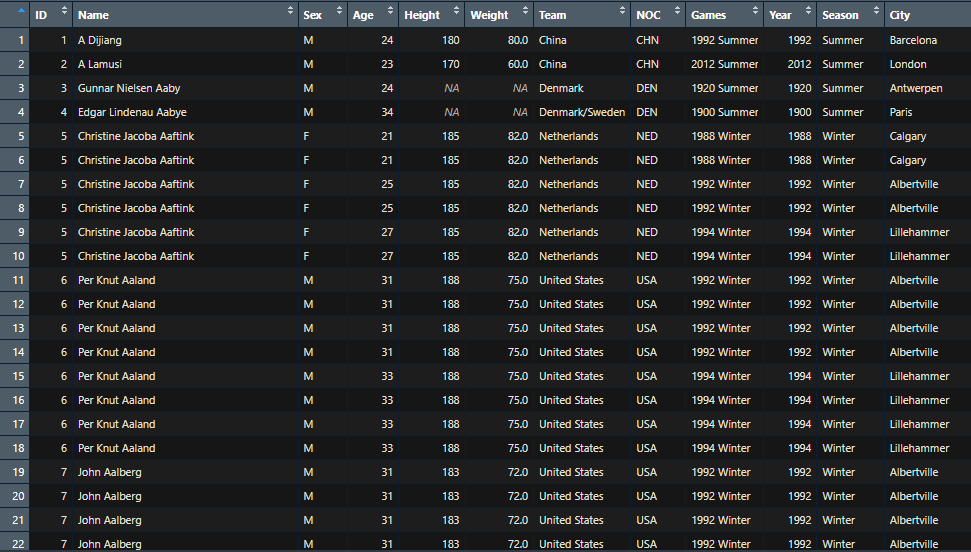


Рисунок 1 – Исходная таблица с данными

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 (гипотезы ) берется утверждение о том, что данные распределены нормально. На рисунке 3 изображен результат работы теста.

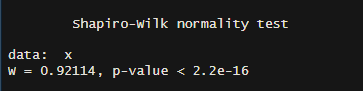


Рисунок 3 – Результат теста Шапиро-Уилкса

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.

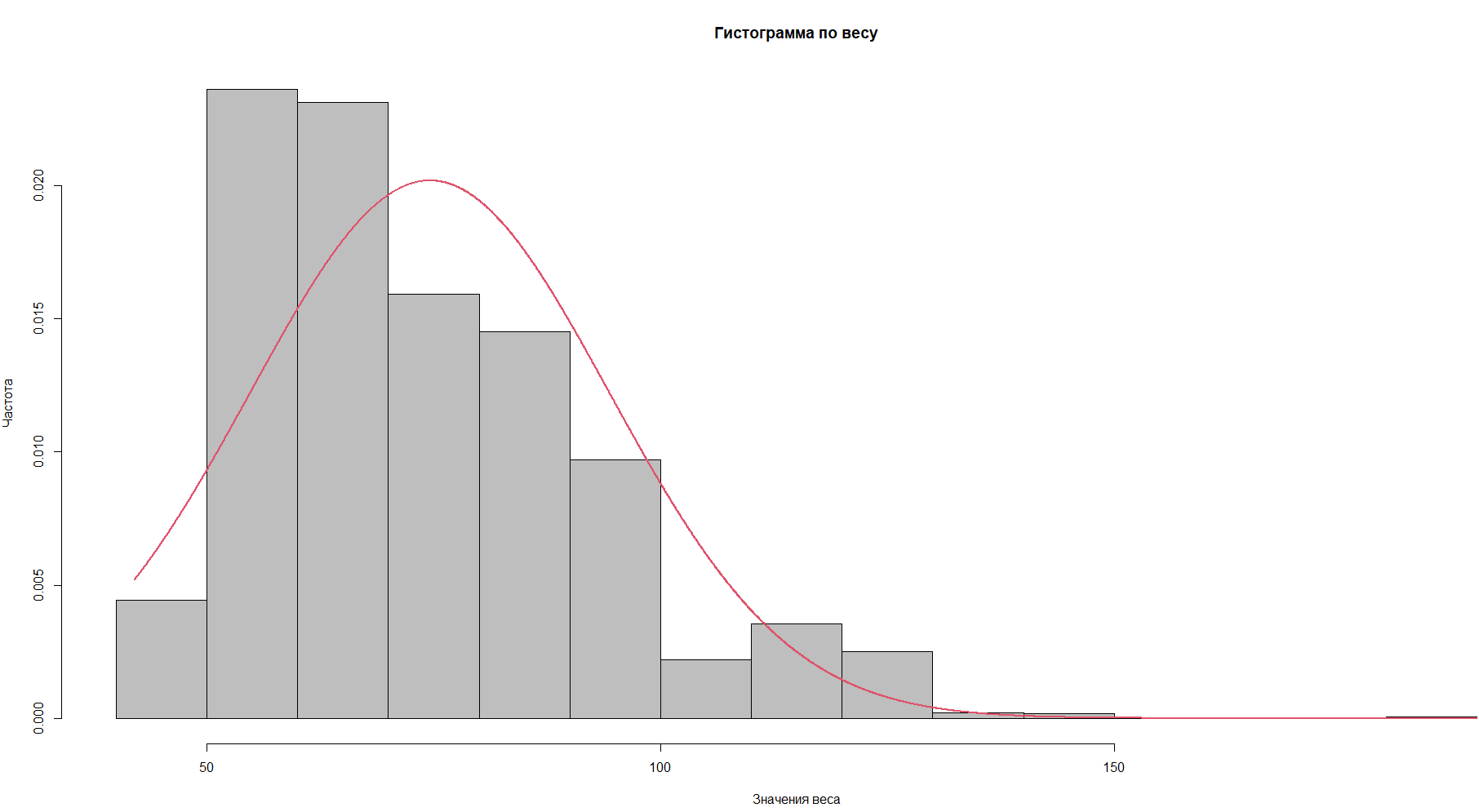


Рисунок 4 – Гистограмма с линией плотности

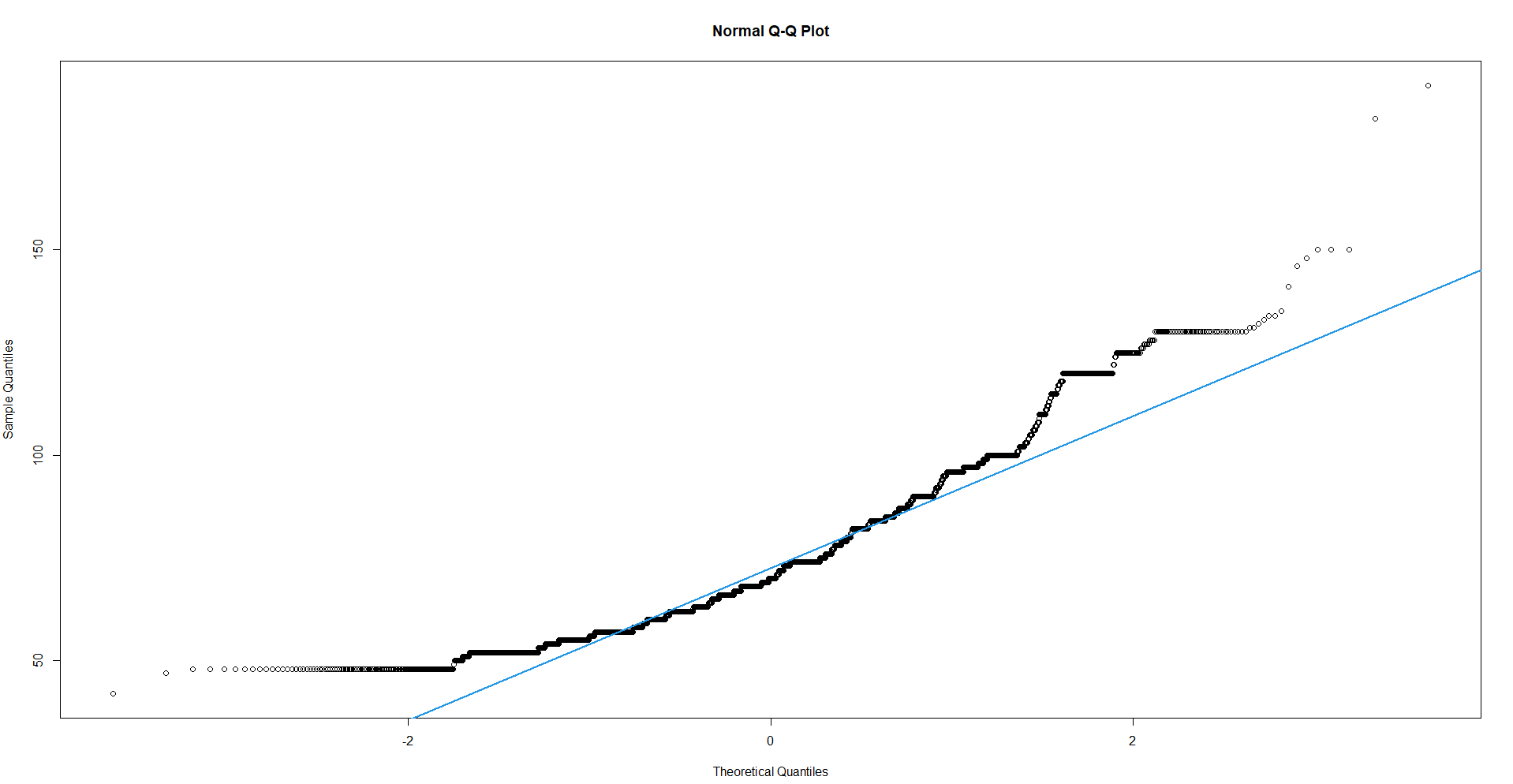


Рисунок 5 – Квантильно-квантильный график

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.

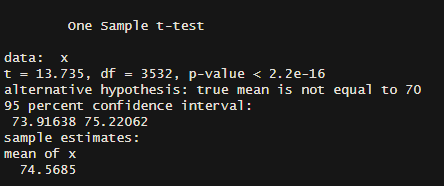


Рисунок 6 – Результат выполнения теста Стьюдента

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.

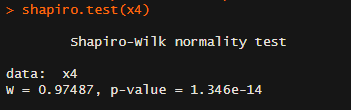


Рисунок 7 – Результат выполнения теста Шапиро-Уилкса

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.

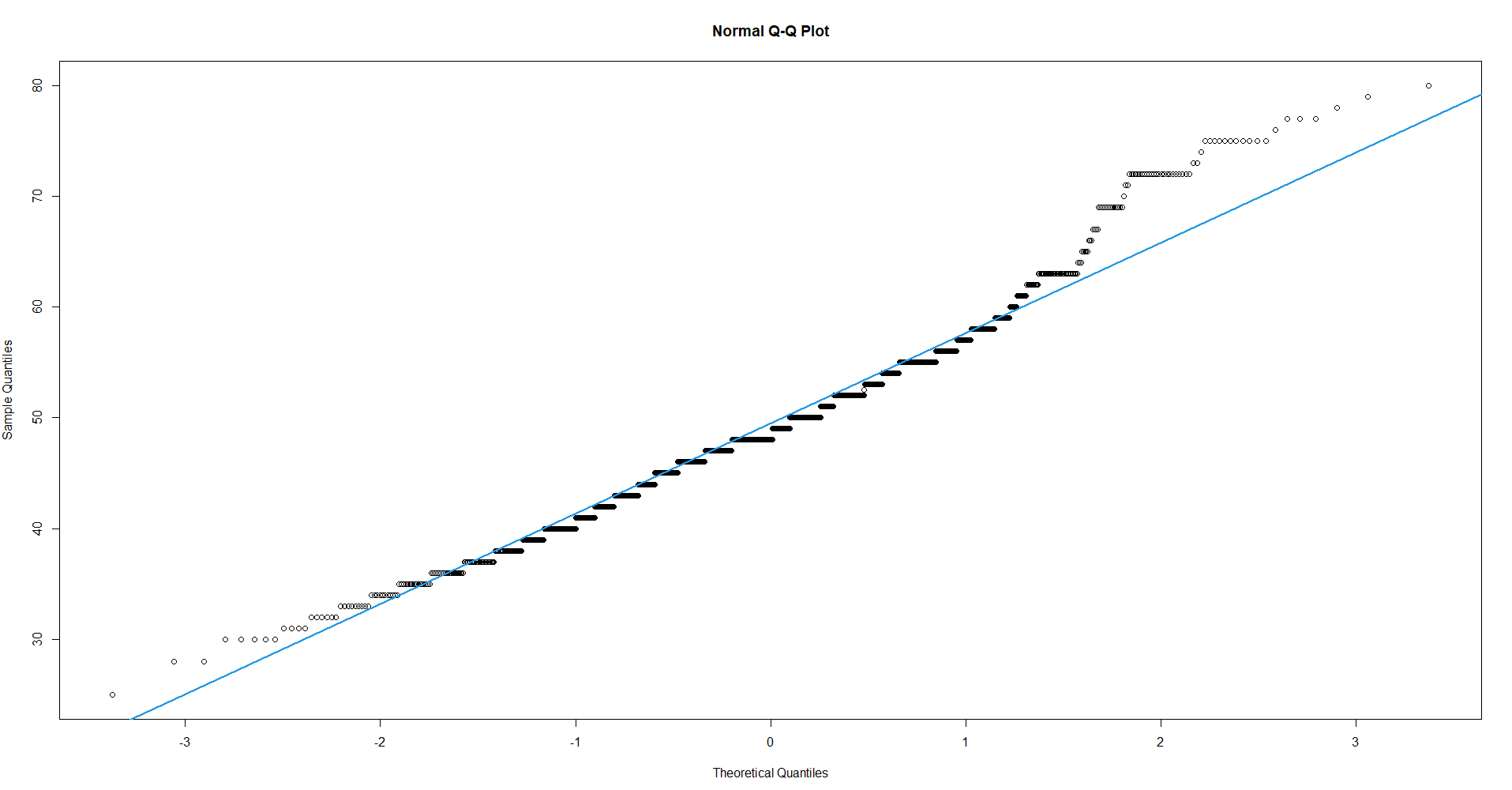


Рисунок 8 – Квантильно-квантильный график для двух видов спорта

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.

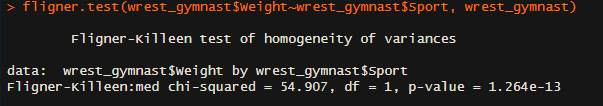


Рисунок 9 – Результаты выполнения теста Флингера-Киллина

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.

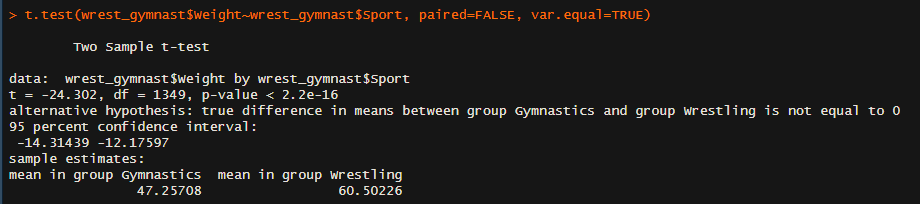


Рисунок 10 – Результат работы t-test

Явно видно, что 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:** В данной работе были ознакомлены с некоторыми статистическими тестами, принципами их работы. Помимо этого, было изучено оценивание нормальности распределения выборки, а также выполнение оценки статистических гипотез.