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 №6.**

**Дисциплина: Обработка больших данных**

Работу выполнил:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Лотарев С.Ю.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Т. А. Приходько

Краснодар

2025

**Тема**: Проверка статистических гипотез.

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

Вариант 9:

|  |  |
| --- | --- |
| Лотарев Сергей Юрьевич | Конькобежный спорт |

**Ход работы**

1. Рассмотрим фрагмент исходного датасета:

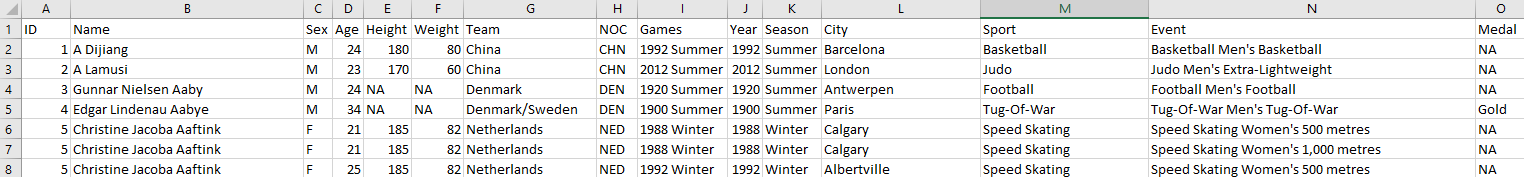


Рисунок 1 – Фрагмент датасета

Датасет содержит данные о достижениях олимпийских спортсменов на олимпийских играх с 1896 по 2016 год. Он включает в себя следующие параметры:

- ID – уникальный номер спортсмена

- Name – имя

- Sex – пол

- Age – возраст

- Height – рост в см

- Weight – вес в кг

- Team – команда, за которую играл спортсмен

- NOC – 3-буквенный код команды, определяемый Национальным олимпийским комитетом

- Games – год и сезон олимпийских игр

- Year – год проведения игр

- Season – сезон проведения игр (летние/зимние)

- City – город проведения игр

- Event – формат соревнования

- Medal – полученная медаль

В рамках лабораторной работы оцениваются данные спортсменов, поэтому ограничимся числовыми параметрами, имеющими прямое отношение к атлетам – имя, пол, возраст, рост и вес.

2. На основе выбранных параметров проведём дескриптивный анализ спортсменов конькобежцев.

Для этого рассмотрим сводку *summary* значений возраста, роста и веса для конькобежцев и всех атлетов:

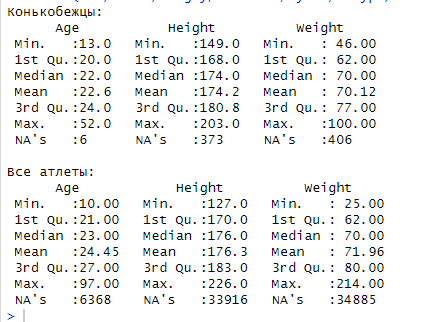
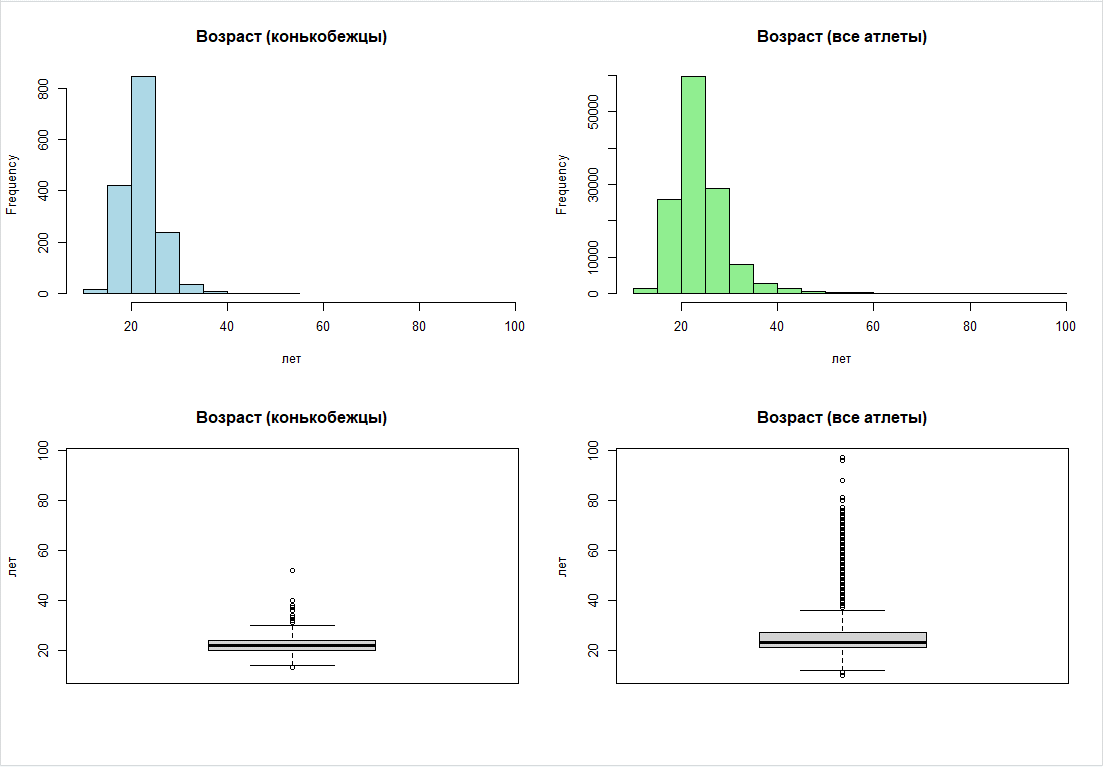


Рисунок 2 – Сводка по параметрам конькобежцев и всех атлетов

а также построим гистограммы и боксплоты по значениям рассматриваемых параметров и проанализируем их.

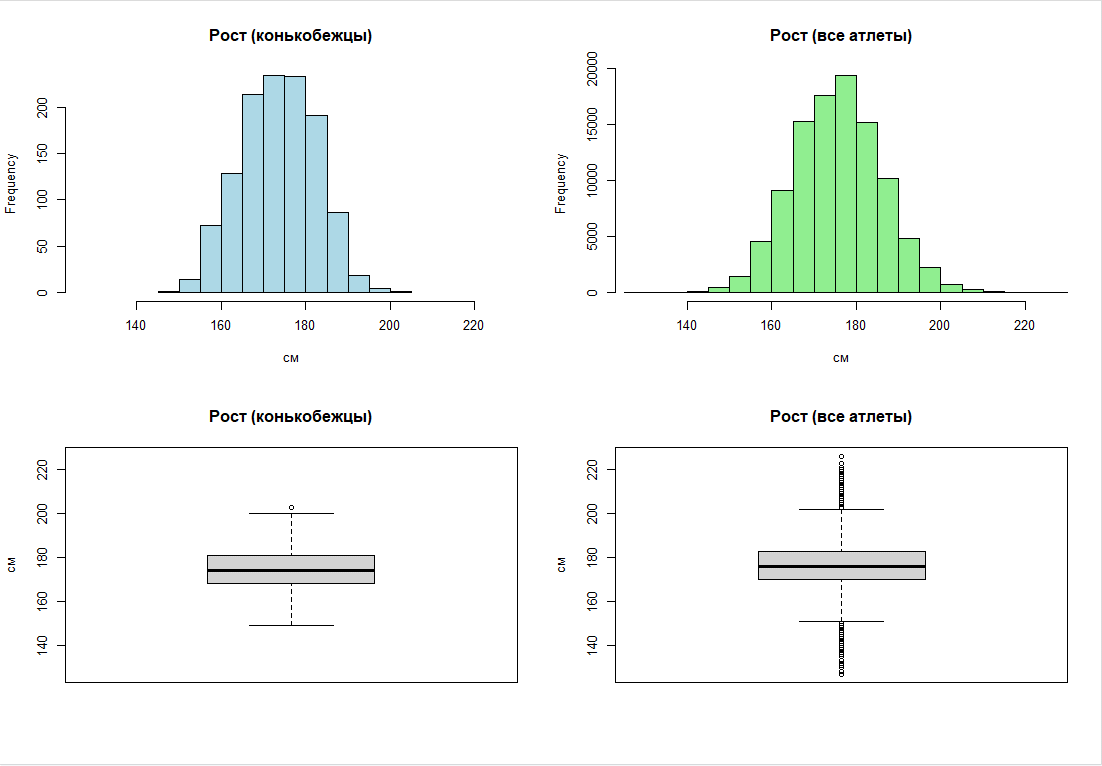
1) Возраст

Рисунок 3 – Графики возраста конькобежцев и всех атлетов

По графикам видно, что возраст конькобежцев распределён более узко по сравнению с общей выборкой всех атлетов. Основная масса спортсменов в этой дисциплине находится в возрасте от 18 до 26 лет, с медианой около 22 лет. Распределение возрастов всех атлетов шире: оно включает как юных участников (до 15 лет), так и возрастных (до 97 лет), с множеством выбросов. Это связано с тем, что в других видах спорта (например, стрельба, конный спорт) участвуют более взрослые спортсмены.

Таким образом, конькобежный спорт предъявляет более строгие возрастные требования — возможно, из-за высокой физической нагрузки и скорости, требующей пика формы в относительно молодом возрасте.

2) Рост

Рисунок 4 – Графики роста конькобежцев и всех атлетов

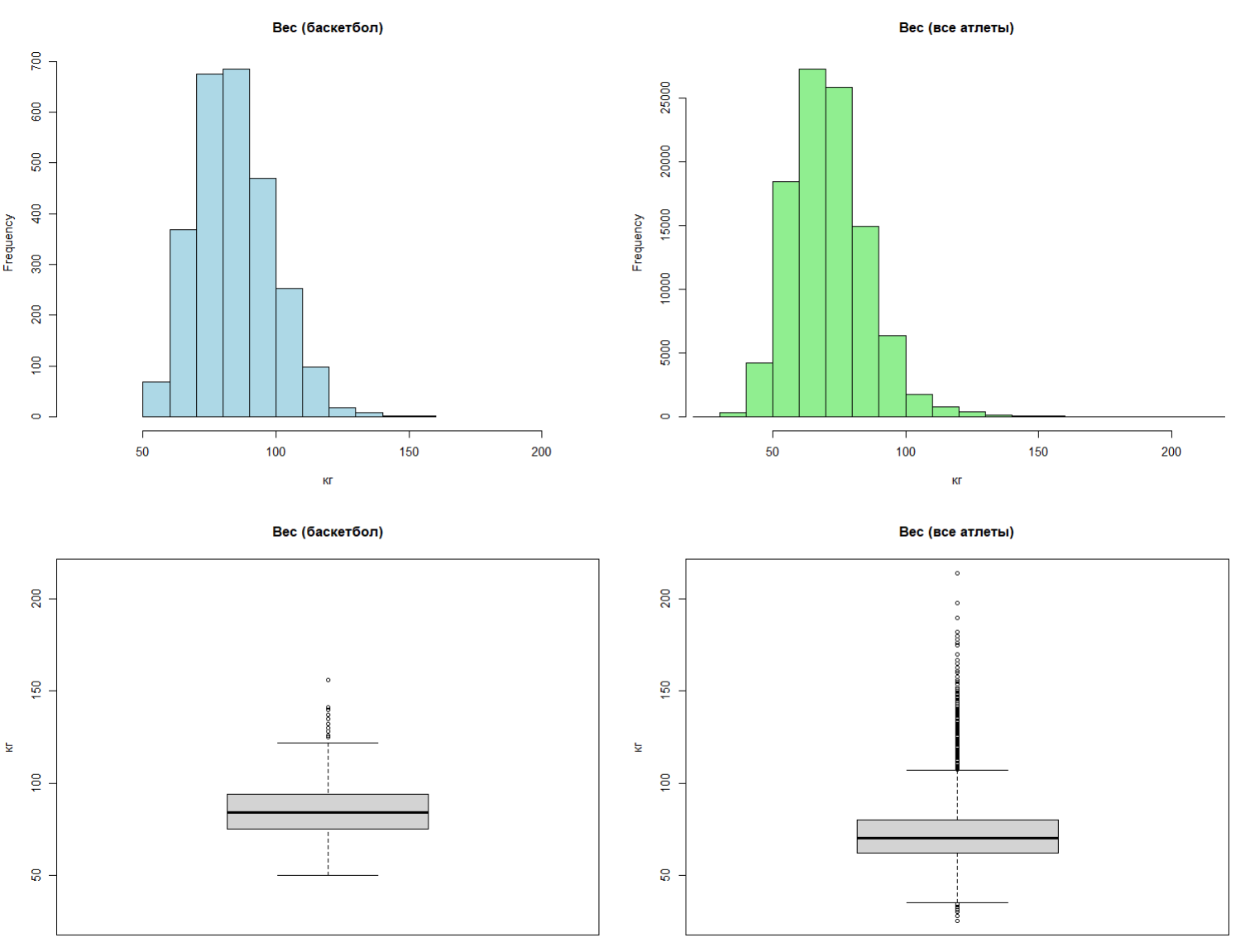
Как видно из графиков, конькобежцы в среднем немного ниже остальных атлетов — медиана роста составляет 174 см, что почти совпадает с медианой по всей выборке (176 см).

Однако стоит отметить, что распределение роста конькобежцев более суженное и симметричное, с минимальным числом выбросов. Основной ростовой диапазон варьируется от 168 до 181 см.

У всей выборки спортсменов диапазон шире — встречаются как очень низкие (до 127 см), так и очень высокие (до 226 см) спортсмены, что отражает разнообразие требований к телосложению в разных видах спорта.

Таким образом, рост у конькобежцев более стандартизирован, что может быть связано с требованиями аэродинамики и мощных длинных рывков в технике скольжения.

3) Вес

Рисунок 5 – Графики веса конькобежцев и всех атлетов

У конькобежцев вес в среднем составляет 70 кг, и распределение сужено: большинство находятся в пределах от 67 до 77 кг, с редкими выбросами. У всех атлетов распределение шире, встречаются значения до 214 кг, много выбросов. Это говорит о более стабильной массе тела в конькобежном спорте, где важны баланс между мышечной массой и аэродинамикой.

3. Проверим выборку веса конькобежцев на нормальность при помощи метода Шапиро-Уилка.

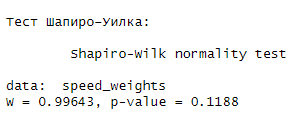


Рисунок 6 – Результат выполнения метода Шапиро-Уилка для выборки конькобежцев

Согласно тесту Шапиро–Уилка, распределение веса конькобежцев не отличается от нормального (W = 0.996, p-value = 0.119). Это подтверждает допустимость предположения о нормальности для дальнейших параметрических тестов.

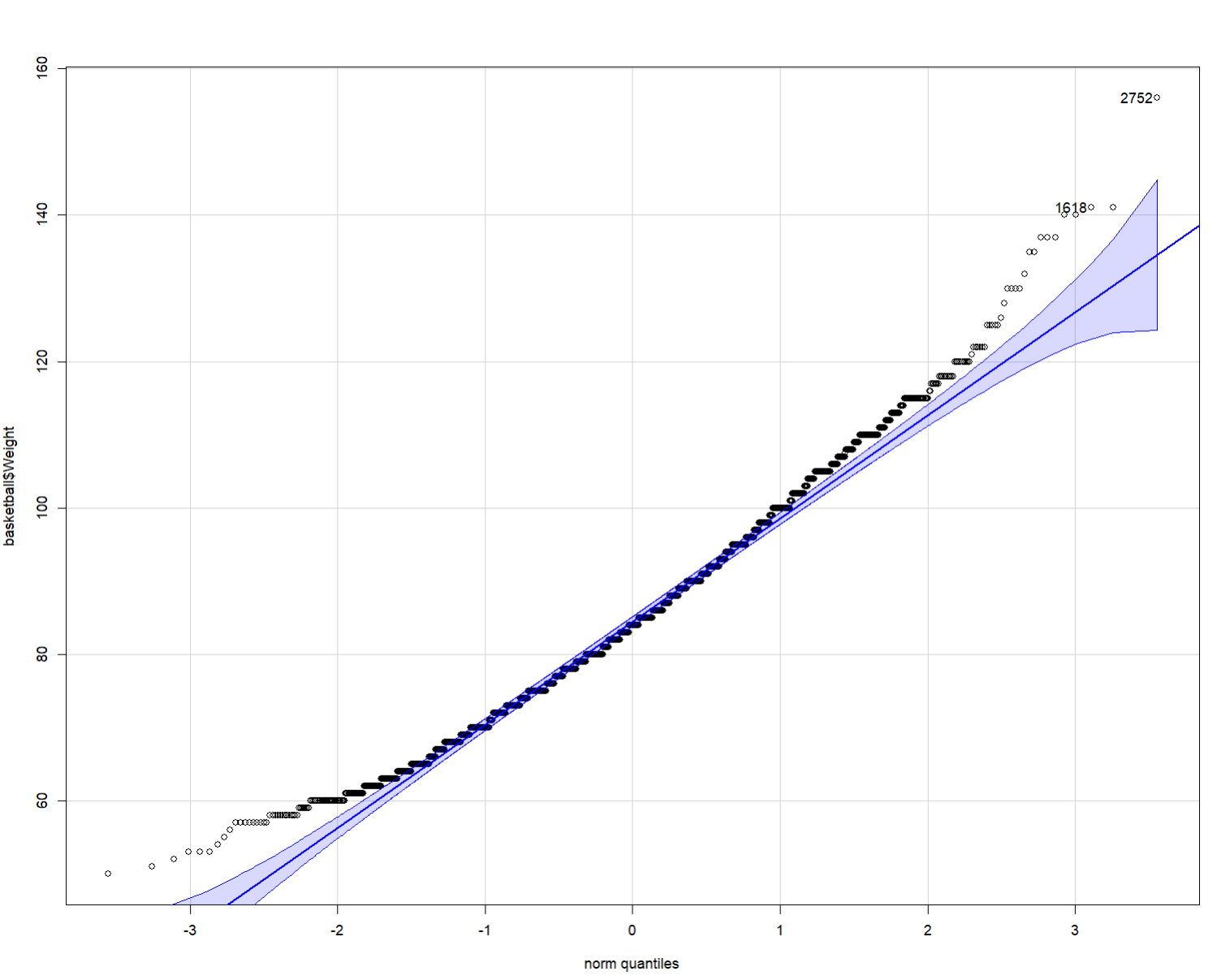


Рисунок 7 – Квантильно-квантильный график выборки веса конькобежцев

Квантильно-квантильный график подтверждает близость распределения веса конькобежцев к нормальному. Точки в основном лежат вдоль диагонали и попадают в доверительную область. Лёгкие отклонения в хвостах не критичны. Это согласуется с результатом теста Шапиро–Уилка (p-value = 0.119).

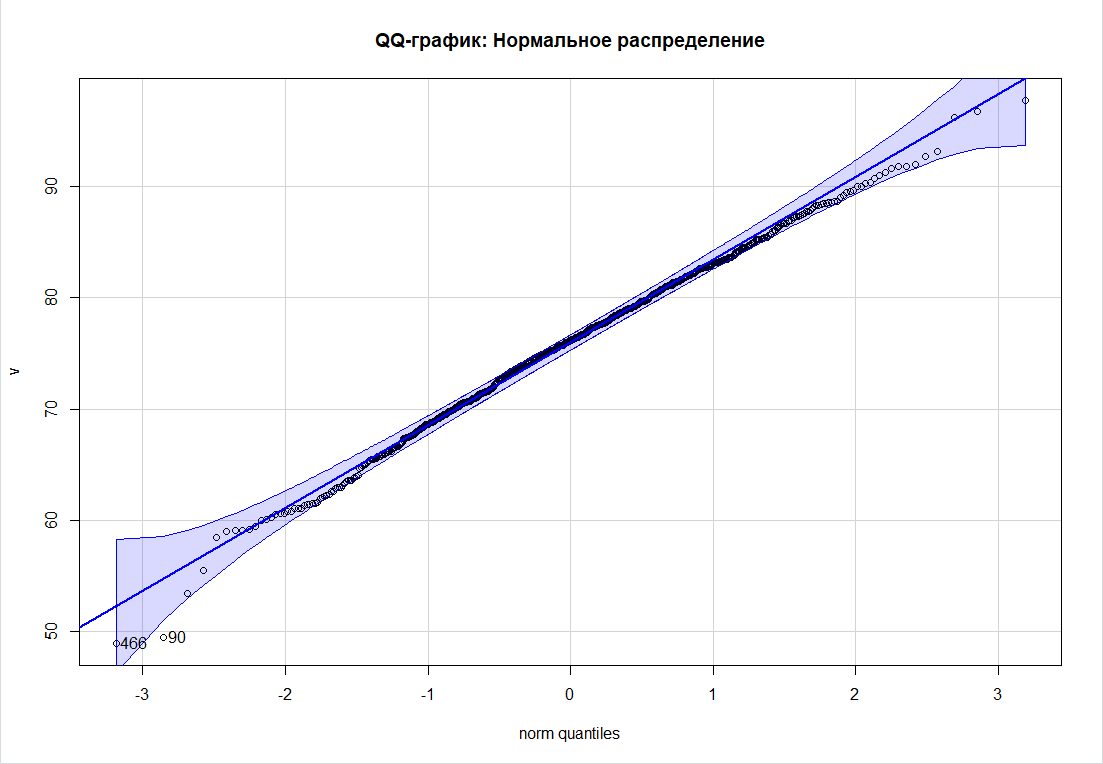


Рисунок 8 – Пример квантильно-квантильного графика для выборки, близкой к нормальному распределению

4. Поскольку, согласно тесту Шапиро–Уилка, распределение веса не отличается от нормального (p-value = 0.119), допускается использование параметрических методов. Тем не менее, для надёжности также применим непараметрический тест Уилкоксона, основанный на медиане, чтобы получить дополнительное подтверждение.

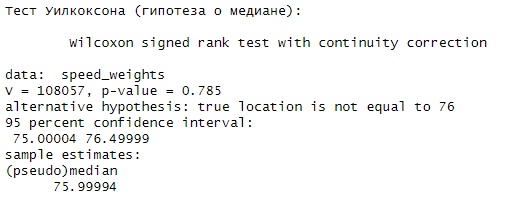


Рисунок 9 – Результаты теста Уилкоксона по гипотезе о среднем весе конькобежцев

По результатам теста Уилкоксона, медиану веса конькобежцев можно принять равной 76 кг (p-value = 0.785). 95% доверительный интервал медианы составляет [75.00; 76.50]. Таким образом, статистически значимых отклонений от предполагаемого значения не выявлено.

Проверим также выполнение теста Стьюдента на гипотезе о среднем арифметическом веса конькобежцев:

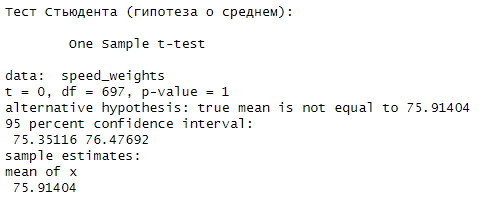


Рисунок 10 – Результаты теста Стьюдента по гипотезе о среднем весе конькобежцев

По результатам теста, среднее значение веса мужчин-конькобежцев составляет 75.91 кг. Поскольку p-value = 1, гипотеза о равенстве среднего указанному значению полностью подтверждается. Доверительный интервал: [75.35; 76.48] кг.

5. Проверим гипотезу о равенстве среднего веса мужчин-конькобежцев и мужчин-фигуристов. Для этого сначала проверим равенство дисперсий с помощью теста Бартлетта.

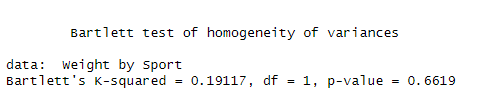


Рисунок 10 – Результаты теста Бартлетта по гипотезе о равенстве дисперсий выборки весов мужчин-конькобежцев и мужчин-фигуристов

Согласно тесту Бартлетта (p-value = 0.662), дисперсии веса у мужчин-конькобежцев и мужчин-фигуристов не различаются статистически значимо. Следовательно, для дальнейшего сравнения средних значений можно использовать классический двухвыборочный t-тест.

6. Перейдём непосредственно к проверке гипотезы.

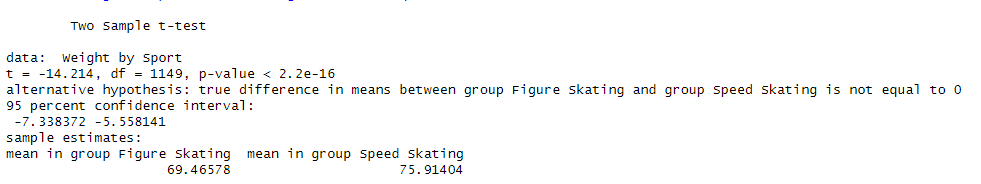


Рисунок 11 – Результаты t-теста по гипотезе о равенстве среднего значения весов мужчин-конькобежцев и мужчин-фигуристов

Двухвыборочный t-тест показал, что средний вес мужчин-конькобежцев (75.91 кг) статистически значимо выше, чем у мужчин-фигуристов (69.47 кг), p-value < 2.2e-16. Разница между средними составляет от **5.6 до 7.3 кг** с доверительной вероятностью 95%.

**Вывод**: в процессе выполнения данной лабораторной работы я изучил статистические тесты, способы оценки нормальности распределения выборки и методы оценки статистических гипотез, а также освоил инструменты языка R для выполнения этих операций. Эти знания я применил для проведения проверки статистических гипотез на основе датасета с данными об олимпийских достижениях атлетов в период с 1896 по 2016 год.