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:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Вавакин В.О.

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 3:

|  |  |
| --- | --- |
| Вавакин Владислав Олегович | Баскетбол |

**Ход работы**

1. Рассмотрим фрагмент исходного датасета:

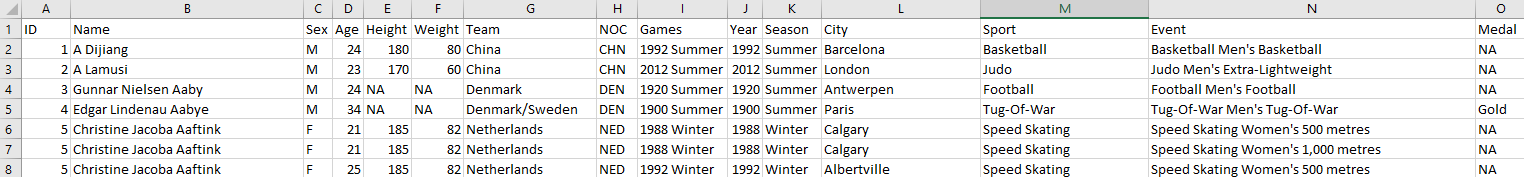


Рисунок 1 – Фрагмент датасета

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-баскетболистов.

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

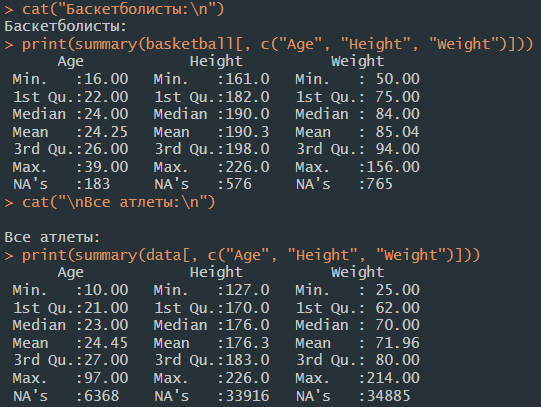
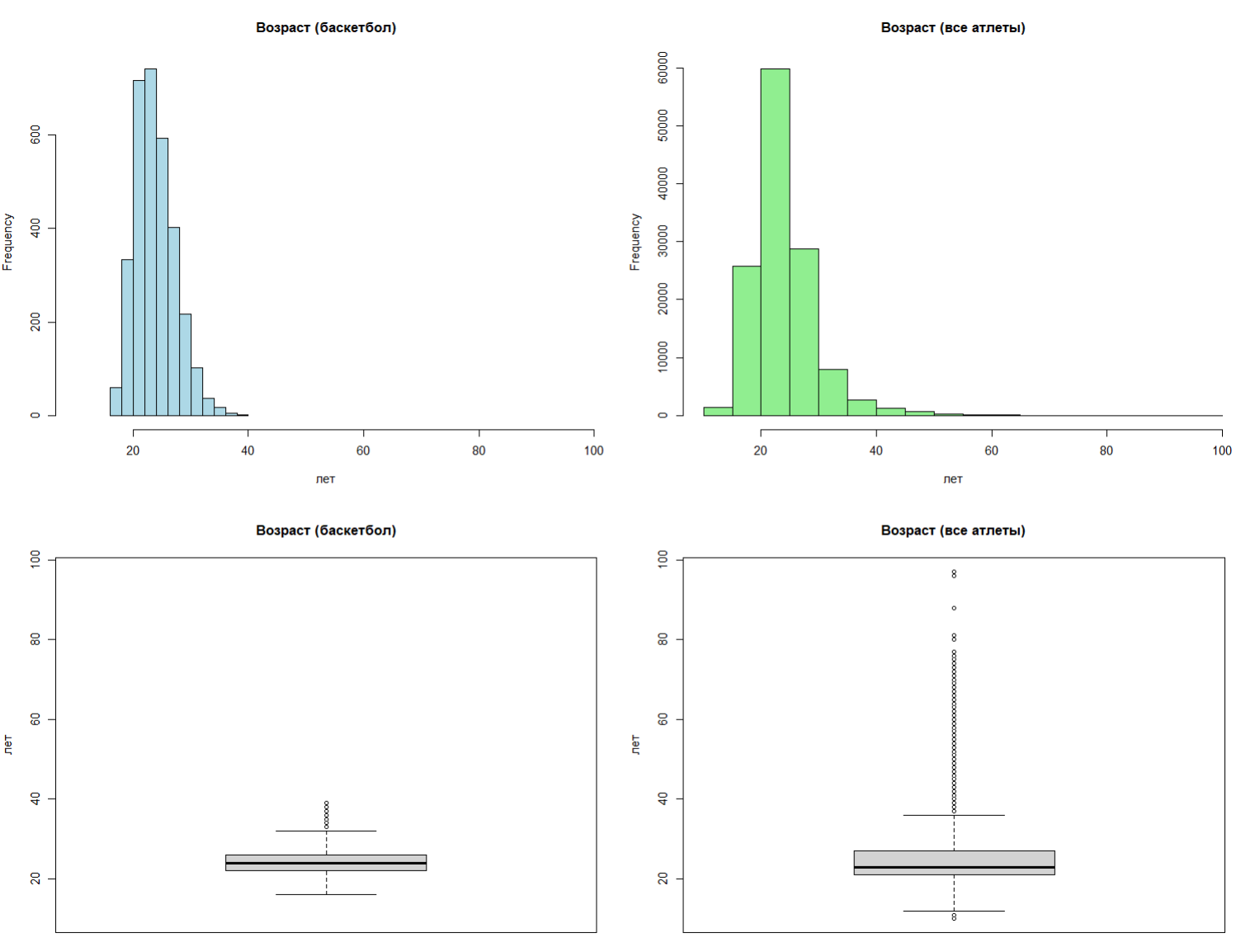


Рисунок 2 – Сводка по параметрам баскетболистов и всех атлетов

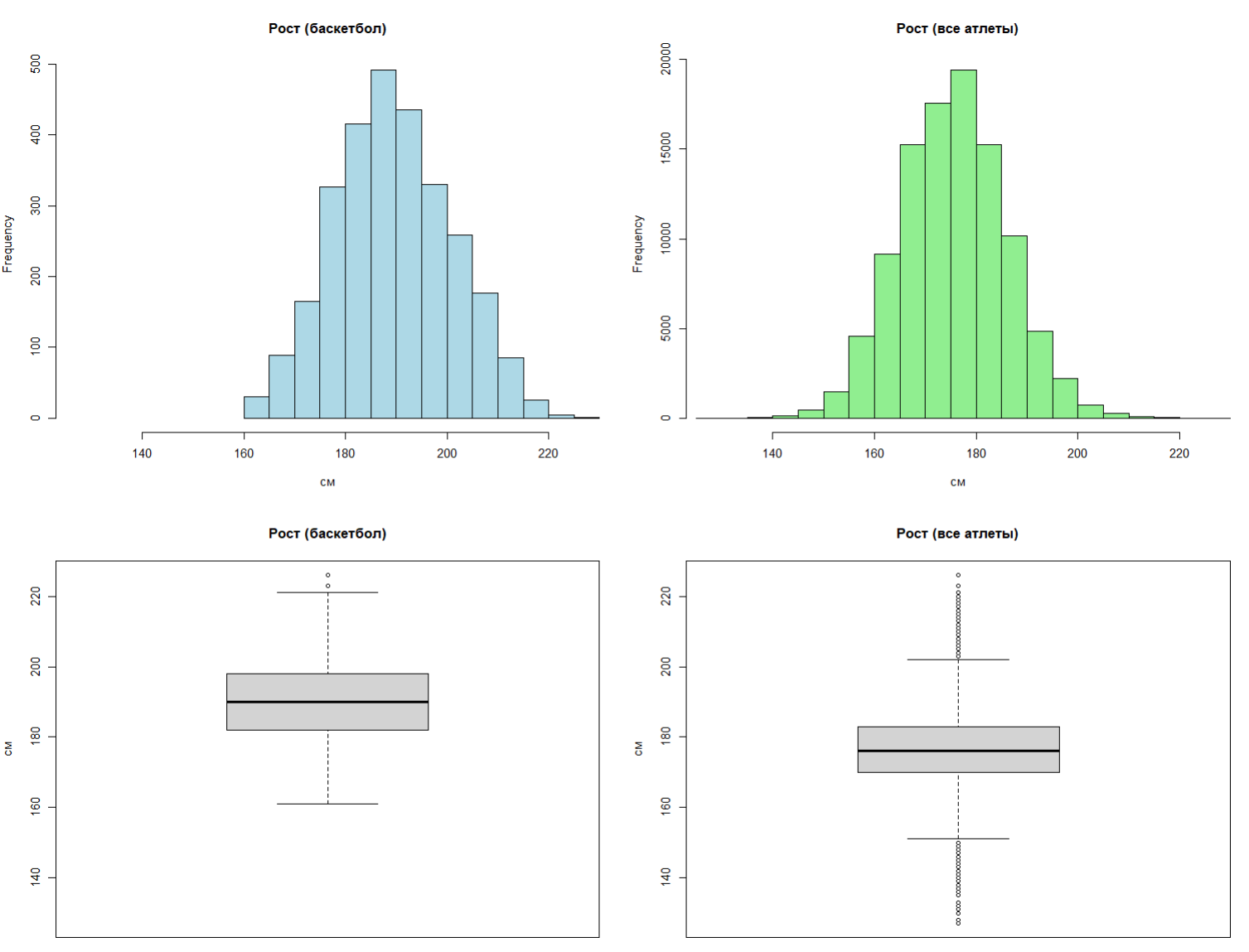
а также построим гистограммы и боксплоты по значениям рассматриваемых параметров и проанализируем их.

1) Возраст

Рисунок 3 – Графики возраста баскетболистов и всех атлетов

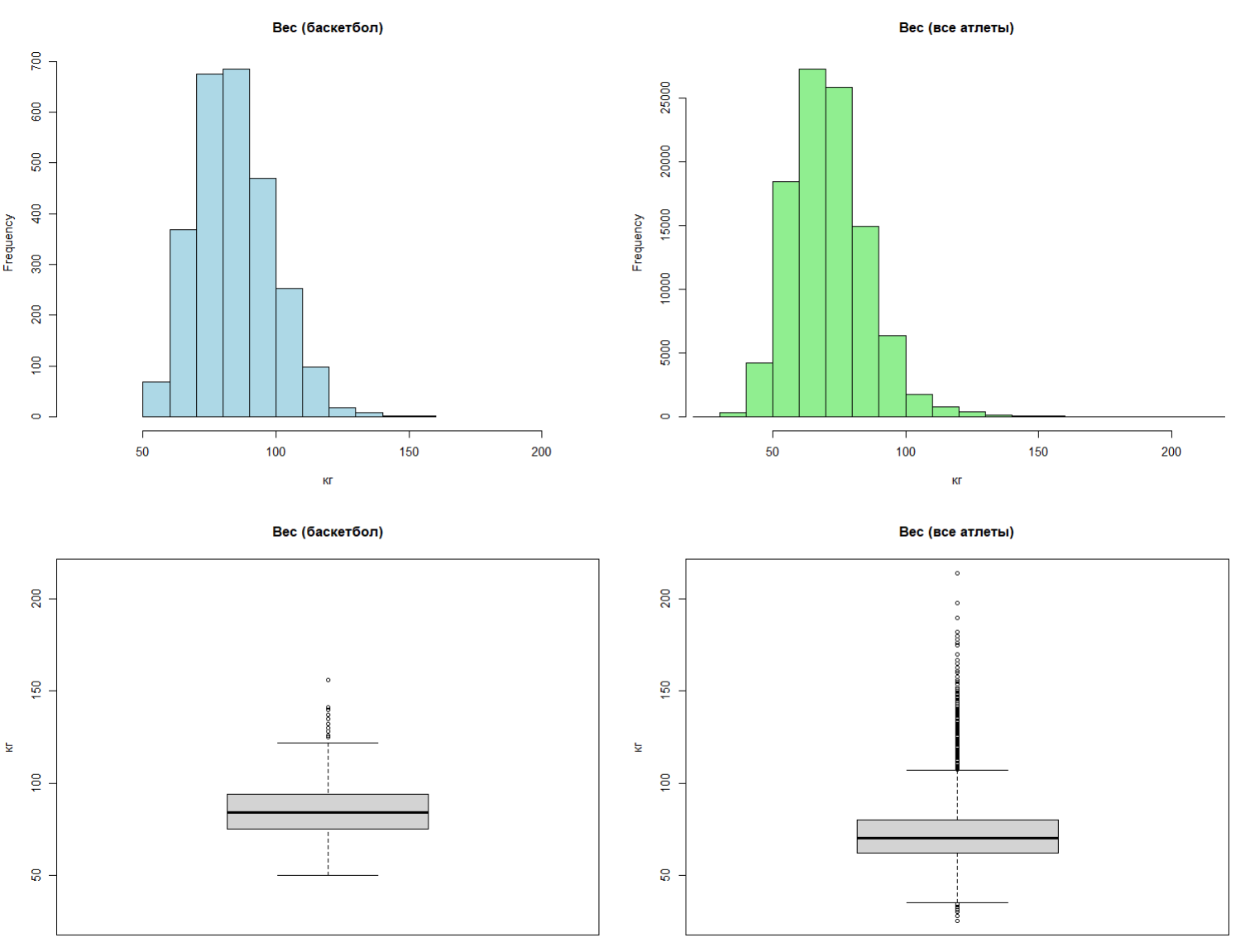
Здесь можно заметить, что возраст баскетболистов в среднем практически не отклоняется от общей выборки. Кроме того, диапазон возрастов баскетболистов достаточно небольшой, и в нём практически отсутствуют выбросы, в отличие от общей выборки, где возраст сильно разнится. Это подтверждается приведённой выше сводкой, откуда можно узнать, что максимальный возраст баскетболистов в датасете – 39 лет, а спортсменов в целом – 97 лет.

2) Рост

Рисунок 4 – Графики роста баскетболистов и всех атлетов

На данных графиках заметно, что баскетболисты в среднем имеют больший рост, чем все атлеты в целом. Из сводки и боксплотов следует, что значение 3-го квартиля роста всех атлетов почти совпадает со значением 1-го квартиля роста баскетболистов. Другими словами, ~75% баскетболистов выше ростом, чем ~75% всех атлетов.

3) Вес

Рисунок 5 – Графики веса баскетболистов и всех атлетов

Баскетболисты также в среднем имеют больший вес по сравнению с общей выборкой. И медиана, и среднее арифметическое этого параметра у баскетболистов превышает значения у всех атлетов приблизительно на 15 кг.

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

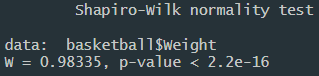


Рисунок 6 – Результат выполнения метода Шапиро-Уилка для выборки баскетболистов

Статистика теста Шапиро-Уилка W равна 0,98, что говорит о высокой визуальной схожести с нормальным распределением, но вероятность p нормальности распределения приблизительно равна 0, что связано с наличием выбросов и отклонения среднего значения выборки. Об этом же говорит квантильно-квантильный график:

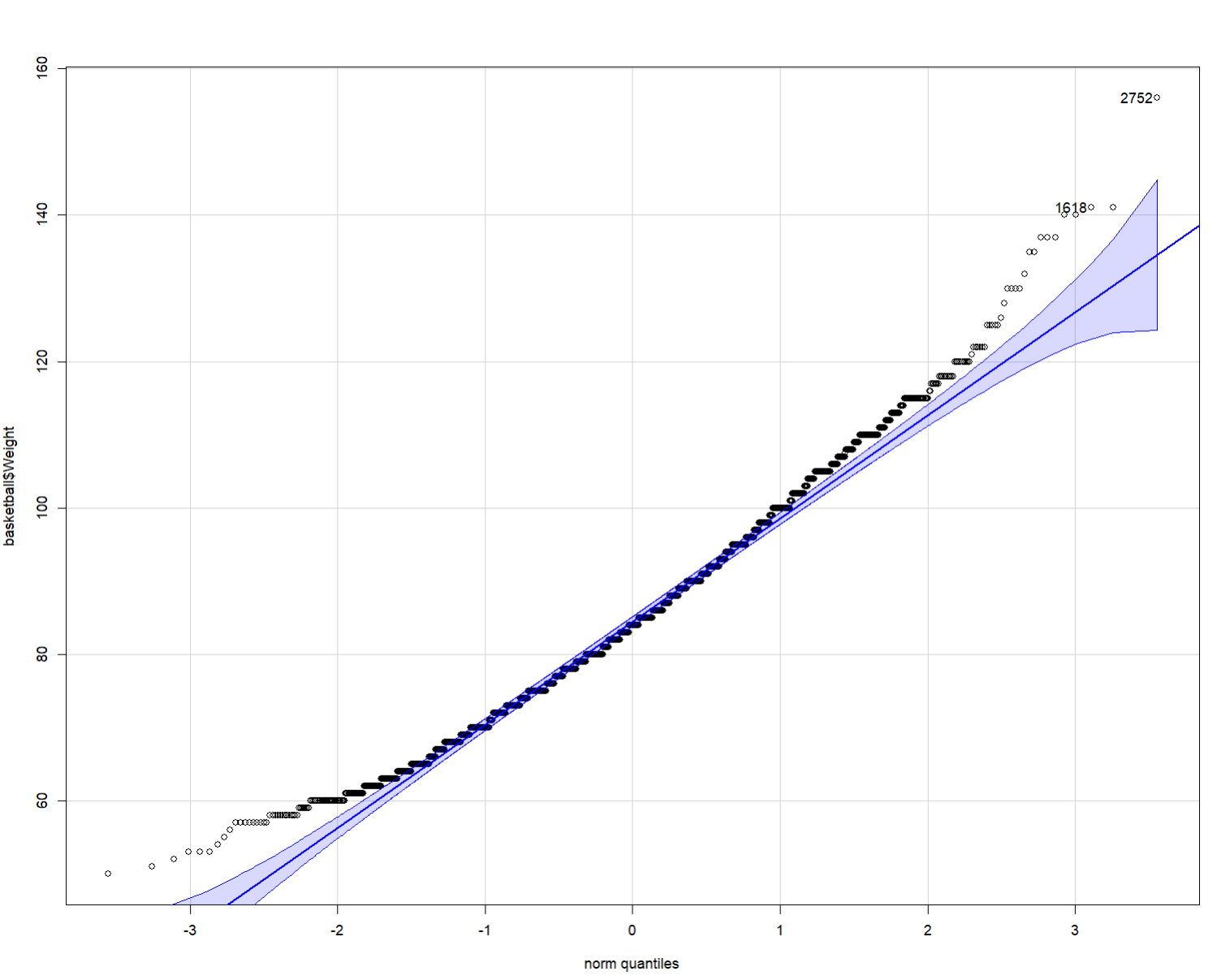


Рисунок 7 – Квантильно-квантильный график выборки веса баскетболистов

На графике видно, что значения на границах диапазона отклоняются от ожидаемых для нормального распределения. Для сравнения график значений того же объёма, сгенерированных функцией *rnorm*, имеет вид:

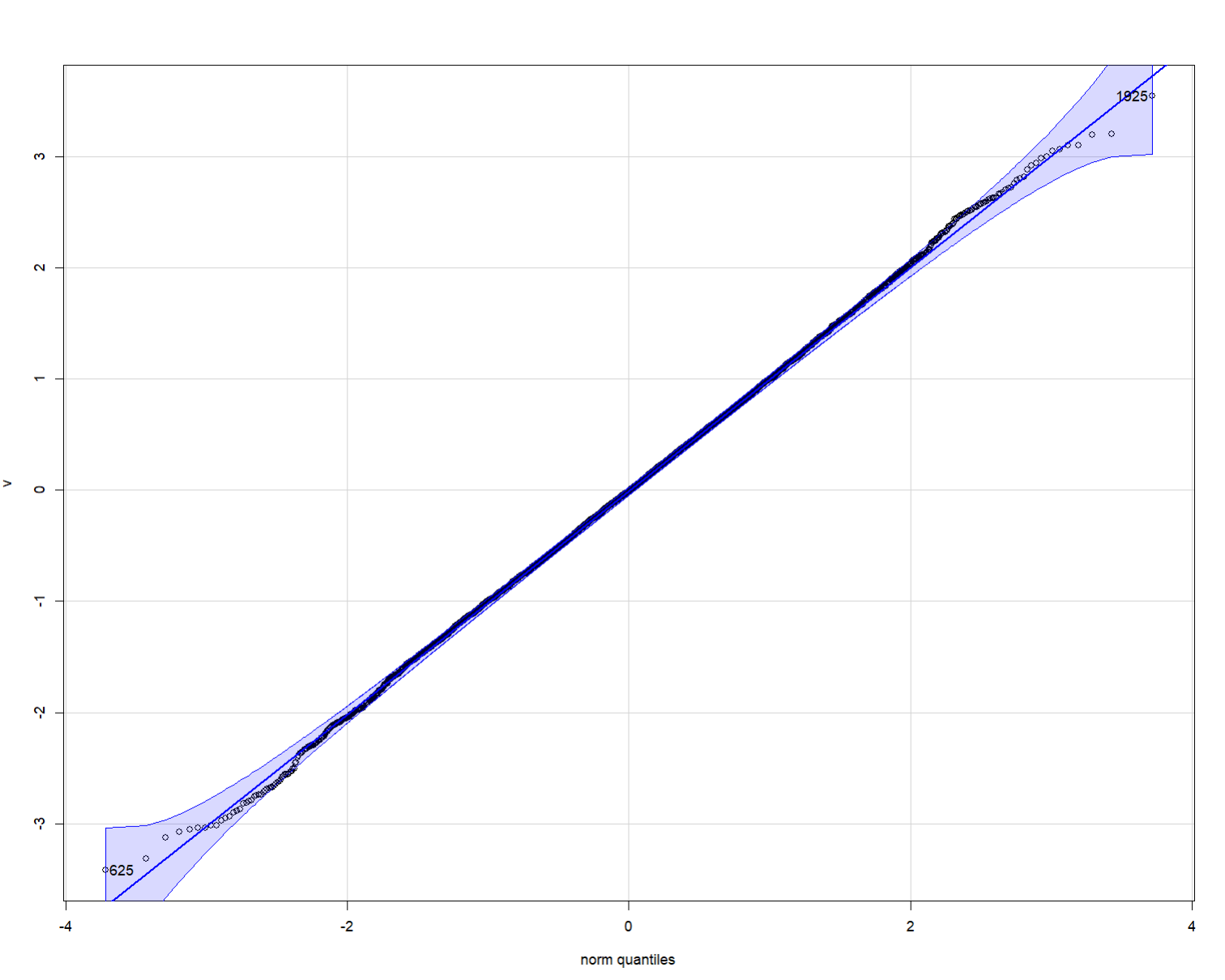


Рисунок 8 – Пример квантильно-квантильного графика для выборки, близкой к нормальному распределению

4. Поскольку, согласно тесту Шапиро-Уилка, выборка мало схожа с нормальным распределением, для проверки гипотезы о среднем весе баскетболистов будем использовать тест Уилкоксона, связанного с медианой и не зависящего от предположения о нормальном распределении.

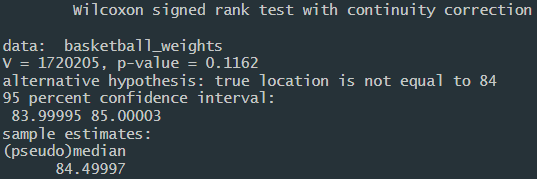


Рисунок 9 – Результаты теста Уилкоксона по гипотезе о среднем весе баскетболистов

Из результатов теста получаем следующие выводы:

- Значение выше порога уровня значимости нулевая гипотеза принимается, истинную медиану можно принять равной 84.

- Вероятность получить значения выборки с истинной медианой 84 кг составляет 11,6%.

- С вероятностью 95% истинная медиана находится в интервале [84, 85] кг.

- Предполагаемая истинная медиана равна ~84,5%.

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

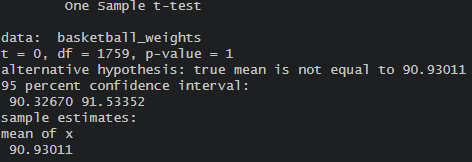


Рисунок 10 – Результаты теста Стьюдента по гипотезе о среднем весе баскетболистов

По результатам теста начальная гипотеза принимается, вероятность справедливости гипотезы, полученная тестом, равна единице, а предположительный интервал равен , однако результаты данного теста нельзя считать надёжными, поскольку предположение о нормальности распределения, необходимое для него, не выполняется.

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

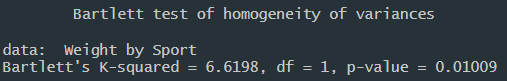


Рисунок 10 – Результаты теста Бартлетта по гипотезе о равенстве дисперсий выборки весов мужчин-баскетболистов и мужчин-рэгбистов

Значение p достаточно мало, гипотезу о равенстве дисперсий можно принять лишь при пороге уровня значимости <0,01.

6. Перейдём непосредственно к проверке гипотезы. Для этого воспользуемся двухвыборочным t-тестом в модификации Уэлча, поскольку дисперсии сравниваемых совокупностей не равны.

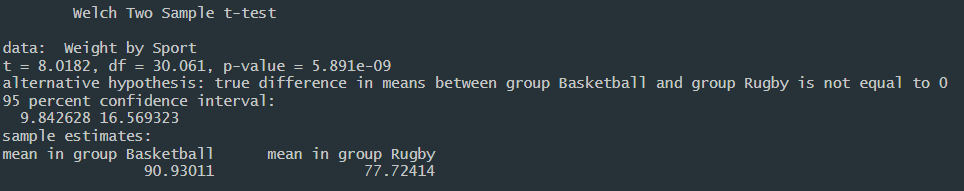


Рисунок 11 – Результаты t-теста по гипотезе о равенстве среднего значения весов мужчин-баскетболистов и мужчин-рэгбистов

Как видно из результатов, средние значения весов у мужчин-баскетболистов и мужчин-рэгбистов статистически значимо различаются. При этом истинная разница между средними значениями с вероятностью 95% находится в диапазоне от 9,84 до 16,57 кг, а ожидаемое истинное среднее значение веса у баскетболистов равно 90,93 кг, у рэгбистов – 77,72 кг.

**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.