Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования «Кубанский государственный университет»

Кафедра вычислительных технологий

**ОТЧЕТ**

о выполнении лабораторной работы № 6

по дисциплине «Обработка больших данных»

Тема: Проверка статистических гипотез

Выполнил: ст. гр. 36/2

Агаджанян А. С.

Проверил: преп.

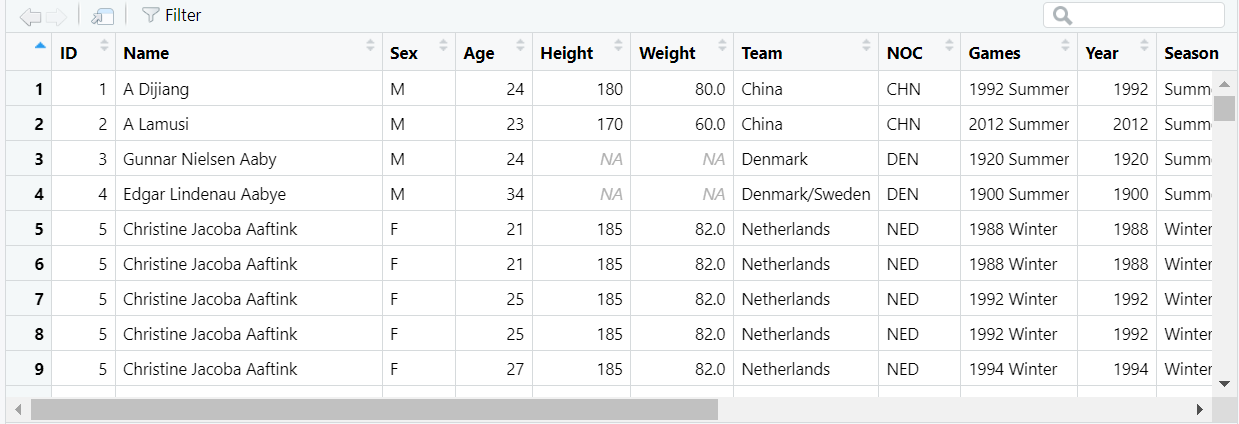
Яхонтов А. А.

Краснодар

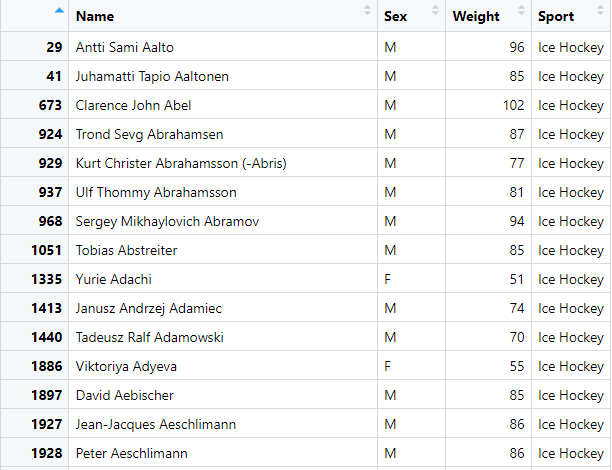
2025

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

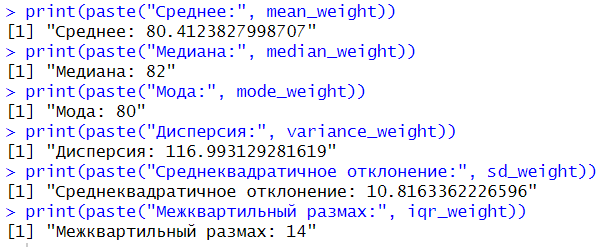
Из исходного csv-файла были импортированы данные в RStudio. Вот часть таблицы:



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 (хоккей на льду), повторяющиеся строчки, а также строчки с пустым значением поля «Вес». Ниже изображена получившаяся таблица.



Был проведен дескриптивный анализ, результаты которого представлены ниже:



Изображение выглядит как текст, диаграмма, снимок экрана, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, снимок экрана, диаграмма, Прямоугольник

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как диаграмма, текст, снимок экрана, График

Контент, сгенерированный ИИ, может содержать ошибки.

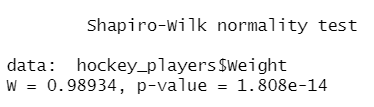
Изображение выглядит как текст, диаграмма, снимок экрана, График

Контент, сгенерированный ИИ, может содержать ошибки.

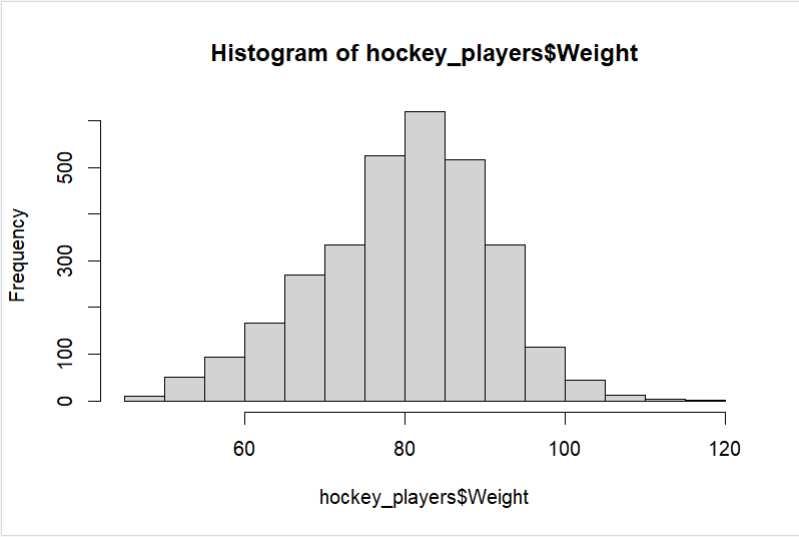
Максимальный шанс (вероятность) допустить ошибку 1-го рода обозначается α (альфа). Это уровень значимости критерия; нулевую гипотезу отвергают, если наше значение p ниже уровня значимости, т. е., если p < α.

Следует принять решение относительно значения α прежде, чем будут собраны данные; обычно назначают условное значение 0,05, хотя можно выбрать более ограничивающее значение, например 0,01. Шанс допустить ошибку 1-го рода никогда не превысит выбранного уровня значимости, скажем α = 0,05, так как нулевую гипотезу отвергают только тогда, когда p < 0,05. Если обнаружено, что p > 0,05, то нулевую гипотезу не отвергнут и, следовательно, не допустят ошибки 1-го рода.

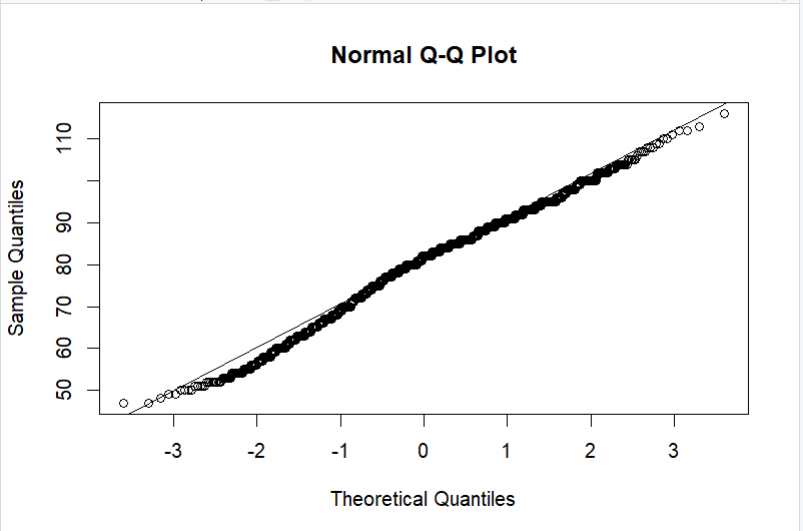
Проведем тест Шапиро-Уилкса на нормальность распределения данных в полученной таблице. В качестве нулевой гипотезы (гипотезы ) берется утверждение о том, что данные распределены нормально. На рисунке 3 изображен результат работы теста.



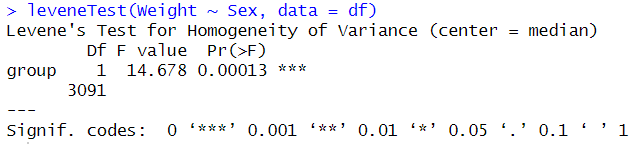
Значение p-value = . Так как это значение меньше, чем граничное значение 0,05, выборка данных не имеет нормального распределения на уровне значимости 0,05. Визуальное подтверждение изображено на следующих рисунках.



Квантильно-квантильный график:



Проведем проверку на дисперсию, для этого используем тест Levene:



Если распределение хотя бы одной из выборок существенно отличается от нормального, в качестве центра предлагается использовать медиану.

В остальных случаях, то есть если распределение каждой выборки можно считать нормальным или несущественно отличающимся от нормального, в качестве центра предлагается использовать среднее арифметическое.

Если центром распределения выбрана медиана, то центры сравниваются с помощью критерия Манна – Уитни-Вилкоксона (непараметрическое).

Если центром распределения выбрано среднее арифметическое, то центры сравниваются с помощью одной из версий критерия Стьюдента (параметрическое).

Исходя из того, что данные распределение отличается от нормального, используем непараметрический тест Уилкоксона.

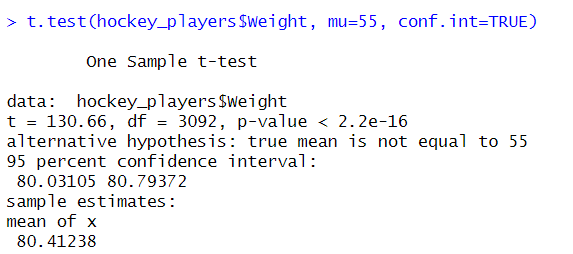
Проверим гипотезу о среднем весе спортсменов, занимающихся хоккеем. Первая гипотеза гласит: «Средний вес спортсменов, занимающихся хоккем, составляет 81 килограмм».

Изображение выглядит как текст, снимок экрана, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Значение p-value = 0.3684 > 0.05, поэтому гипотеза не отвергается. Доверительный интервал расположен от 80.5 до 81.00008, именно в этот интервал входит настоящий средний вес спортсменов – 80,99998.

Тест Стьюдента имеет похожие результаты:



**Сравнение двух независимых выборок**

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

Сперва вновь данные проверяются на нормальность распределения тестом Шапиро-Уилкса.

Изображение выглядит как текст, Шрифт, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

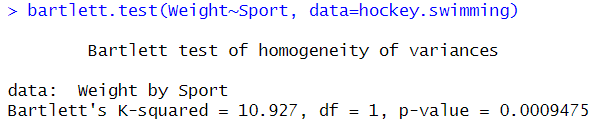
По результатам теста, значение p-value вновь меньше граничного 0.05. Следовательно, гипотеза о нормальном распределении отвергается, квантильно-квантильный график для двух видов спорта представлен ниже:

Изображение выглядит как текст, снимок экрана, График, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Для проверки равенства дисперсий для женщин, занимающихся хоккем и плаванием, можно использовать критерий Бартлетта или тест Флингера-Киллина. Основная гипотеза – дисперсии равны.

Критерий Бартлетта является параметрическим и основан на предположении о нормальности выборок.

А тест Флингера-Киллина является непараметрическим и отлично подходит для анализа данных с ненормальным распределением.

Изображение выглядит как текст, Шрифт, снимок экрана, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Гипотеза отвергнута, дисперсии не равны.

Проверим гипотезу о равенстве среднего веса женщин в двух разных видах спорта. Для этого выполним U-тест Манна-Уитни:

Изображение выглядит как текст, Шрифт, снимок экрана, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Явно видно, что значение p-value близко к нулю, а значит гипотеза не верна и равенство среднего веса спортсменок по гимнастике и плаванию не подтверждено.

Попробуем применить двухвыборочный тест Стьюдента.

Изображение выглядит как текст, Шрифт, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Результаты те же. Этот тест также выводит средний вес, поэтому можно наглядно убедиться, что средний вес спортсменок по хоккею больше на 4 кг веса спортсменок по плаванию.

Вывод: при выполнении данной лабораторной работы ознакомлен с некоторыми статистическими тестами. Помимо этого, было изучено оценивание нормальности распределения выборки, а также выполнение оценки статистических гипотез.