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**

**Отчёт**

**по лабораторной работе №7**

**Дисциплина: ОБРАБОТКА БОЛЬШИХ ДАННЫХ**

**Тема: «Проверка статистических гипотез»**

Работу выполнила \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_А.М. Нагалевский

Направление подготовки 02.03.02 Фундаментальная информатика и

информационные технологии

Направленность (профиль) Математическое и программное обеспечение

компьютерных технологий

Преподаватель

ассистент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В.И.Шиян

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

**Задание:** Набор данных для исследования Аthlen\_Events– это исторический набор данных о современных Олимпийских играх, включая все игры от Афин 1896 года до Рио-2016. Эти данные собраны из www.sports-reference.com в мае 2018 года.

1. Проверьте гипотезу о среднем весе спортсменов выбранного вида спорта.

2. Проверьте гипотезу о равенстве среднего веса женщин (мужчин) в двух разных выбранных видах спорта (сравнение двух независимых выборок – двухвыборочный критерий.).

**Ход работы:**

Из исходного csv файла данные были импортированы в RStduio. Фрагмент таблицы представлен на Рисунке 1.

Изображение выглядит как текст, число, документ, снимок экрана

Автоматически созданное описание

Рисунок 1 – Фрагмент исходной таблицы с данными

Построена новая таблица, содержащая имена, пол и вес спортсменов,

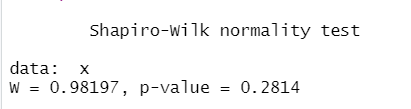
занимающихся горнолыжным спортом.

Изображение выглядит как текст, число, снимок экрана, документ

Автоматически созданное описание

Рисунок 2 – Фрагмент новой таблицы с данными

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.



Изображение выглядит как текст, Шрифт, снимок экрана, белый

Автоматически созданное описание

Рисунок 3 – Результат теста Шапиро-Уилкса

Значение p-value равно 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.

Изображение выглядит как диаграмма, График, линия, текст

Автоматически созданное описание

Рисунок 4 – Гистограмма с линией плотности

Изображение выглядит как текст, диаграмма, График, линия

Автоматически созданное описание

Рисунок 5 – Квантильно-квантильный график

Далее проверим гипотезу о среднем весе спортсменов, занимающихся горнолыжным спортом. В качестве первой гипотезы возьмем заявление о том, что средний вес спортсменов равен 75. Для проверки используем тест Стьюдента.

Результат выполнения теста показан на рисунке 6.

Изображение выглядит как текст, снимок экрана, Шрифт, алгебра

Автоматически созданное описание

Рисунок 6 – Результат выполнения теста Стьюдента

Значение p-value низкое, поэтому нулевую гипотезу нельзя принять. Доверительный интервал расположен от 65.67 до 66.08, именно в этот интервал входит настоящий средний вес спортсменок – 65.88.

Далее необходимо проверить гипотезу о равенстве среднего веса у спортсменок, занимающихся конькобежным спортом и горнолыжным спортом.

Сначала проверим данные на нормальность распределения тестом Шапиро-Уилкса

Изображение выглядит как текст, Шрифт, снимок экрана, белый

Автоматически созданное описание

Рисунок 7 – Результат выполнения теста Шапиро-Уилкса

По результатам теста, значение p-value очень мало, а значит гипотеза о

нормальном распределении не подтвердилась. Для большей наглядности

построим графики для данных по скоростному скейтингу и горнолыжному спорту.

График изображён на рисунке 8.

Изображение выглядит как текст, линия, диаграмма, снимок экрана

Автоматически созданное описание

Рисунок 8 – Квантильно-квантильный график для 2-х видов спорта

Проведём тест на равенство дисперсий. Гипотеза H0 – дисперсии равны – принимается за основную. На рисунке 9 изображён результат теста.

Изображение выглядит как текст, Шрифт, снимок экрана, белый

Автоматически созданное описание

Рисунок 9 – Тест на равенство дисперсий

Значение p-value близко к нулю, поэтому гипотезу можно отвергнуть.

Проверим с помощью t-test гипотезу об отсутствии разницы в среднестатистическом значении среднего веса спортсменок конькобежного спорта и горнолыжного спорта. На рисунке 10 показан результат теста.

Изображение выглядит как текст, Шрифт, снимок экрана, алгебра

Автоматически созданное описание

Рисунок 10 – Результат работы t-test

Значение p-value близко к нулю, а значит гипотеза не верна и равенство среднего веса спортсменок спортсменок конькобежного спорта и горнолыжного спорта не подтверждено. Этот тест выводит также средний вес, поэтому можно наглядно убедиться в правильности результата.