МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«Кубанский государственный университет»**

**(ФГБОУ ВО «КубГУ»)**

**Кафедра вычислительных технологий**

**ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №7**

**Дисциплина: Обработка больших данных**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Минин К.С.

(подпись)

Факультет Компьютерных технологий и прикладной математики

Направление подготовки 02.03.02 Фундаментальная информатика и информационные технологии курс 3

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

Краснодар 2022

**Лабораторная работа № 7**

**«Проверка статистических гипотез»**

**Цель работы:**

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

**Описание данных:** 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. Файл athlete\_events.csv содержит 271116 строк и 15 столбцов. Каждый ряд соответствует отдельному спортсмену, участвующему в индивидуальном Олимпийском соревновании (athlete-events).

**Ход работы:**

Любой вид анализа данных требует (после этапа сбора этих данных) выдвижения некоего предположения (гипотезы) относительно характеристик выбранной совокупности.

**Статистической гипотезой** называют любое предположение о виде или свойствах распределения генеральной совокупности. Мы будем рассматривать две гипотезы: **нулевую (основную) *H0*** и **альтернативную (конкурирующую) *H1***.

**Нулевая гипотеза *H0*** подлежит проверке, по результатам которой ее можно принять либо отклонить. «Принять» означает «не получить убедительных аргументов для отклонения гипотезы».

**Альтернативная гипотеза H1** принимается только тогда, когда есть убедительное статистическое доказательство для отклонения основной гипотезы.

**Ошибкой первого рода** называют ситуацию, в которой отвергается верная гипотеза ***H0***. При **ошибке второго рода** принимается гипотеза ***H0***.в то время, как она неверна.

Считываем таблицы со спортсменами, выбираем вид спорта плавание, отдельно мужчин и отдельно отбираем вес. Построим гистограмму веса для них (рисунок 1).

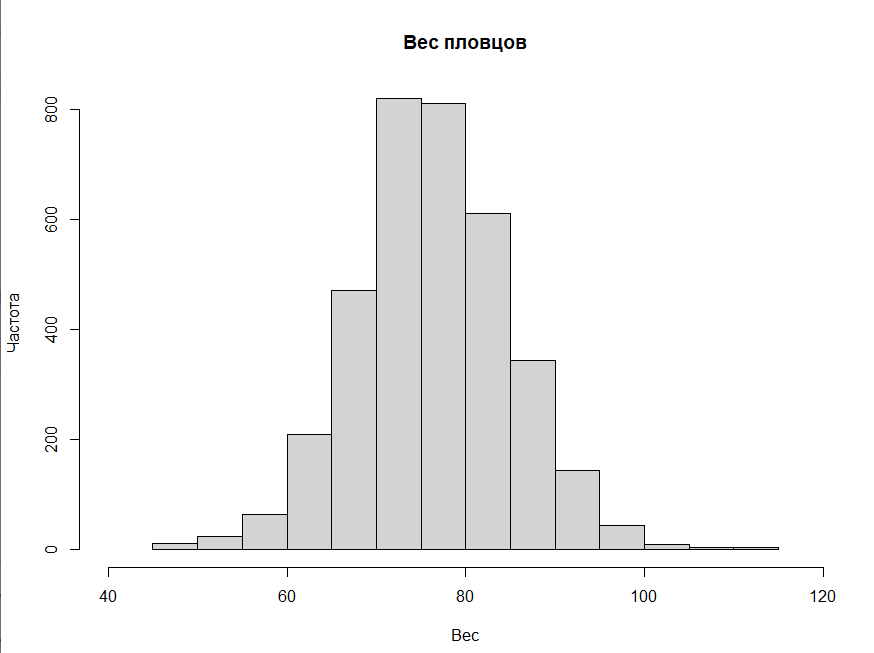
****

Рисунок 1 – Гистограмма веса мужчин в плавании.

Проверим, насколько распределение похоже на нормальное (рисунок 2)

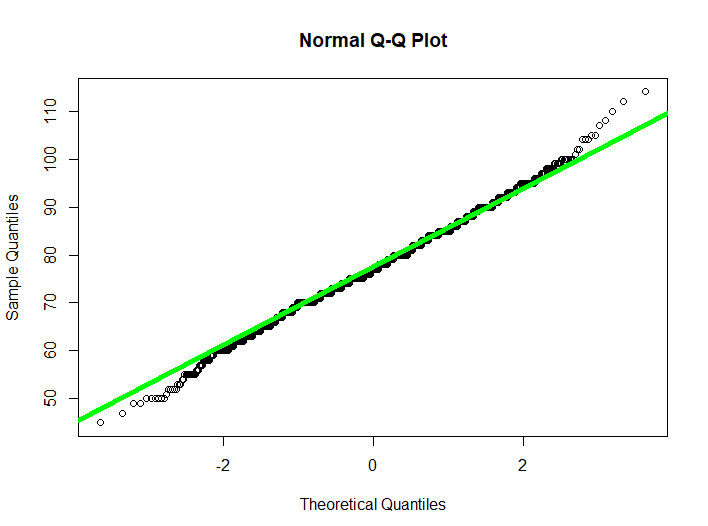


Рисунок 2 – Графическая проверка выборки на нормальность - ответ положительный.

Начнём с одномерных тестов, которые позволяют проверить утверждения относительно того, как распределены исходные данные.

Мной был выдвинута гипотеза о том, что средний вес спортсменов мужского пола, участвовавших в плавании, составляет 78 килограмм. Для проверки был использован тест Стьюдента для одномерных данных. Статистические тесты пытаются высчитать, так называемую, тестовую стати­стику. Затем на основании этой статистики рассчитывается p-величина или p-value, отражающая вероятность ошибки первого рода (ее ещё называют «ложной тревогой»). Принято считать, что нулевой гипотезе соответствует ситуация «по умолчанию». Вычисленная p-величина используется для сравнения с заранее заданным порогом (уровнем) значимости. Если p-величина ниже выбранного порога уровня значимости (как правило 0,05), то нулевая гипотеза отвергается, а если выше, то принимается. В этом примере гипотеза отвергается.

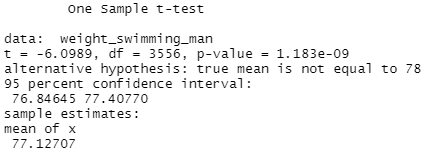


Рисунок 3 – Тест Стьюдента.

Существует так же непараметрический аналог этого теста, то есть теста, не связанного предположениями о нормальном распределении. Это ранговый тест Уилкоксона (Wilcoxon signed-rank test).

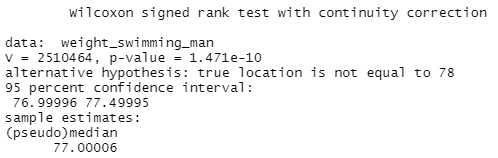


Рисунок 4 – Тест Уилкоксона.

Эта функция выводит практически то же самое. Обратим внимание, что тест связан не со средним, а с **медианой**. Соответственно, вычисляется (если задать conf. int=TRUE) доверительный интервал. Здесь он значительно уже.

Некоторые статистические методы (например, ANOVA или дисперсионный ана­лиз) основаны на том, что данные имеют нормальное распределение. Поэтому вопрос соответствует ли распределение данных нормальному или хотя бы напо­минает нормальное является очень и очень важным. В R реа­лизовано несколько разных техник, отвечающих на вопрос о нормальности. Во-первых, это статистические тесты. Самый простой из них — тест Шапиро-Уилкса (Shapiro-Wilk test):

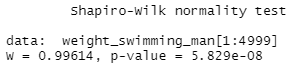


Рисунок 5 – Тест Шапиро-Уилкса.

Таким образом, на основании теста Шапиро-Уилкса можно заключить, что распределение данных в weight существенно отличается от нормального.

Далее рассмотрим сравнение двух независимых выборок для пловцов и спортсменов водного поло мужчин. Проведем проверку на нормальность данных (рисунок 6).

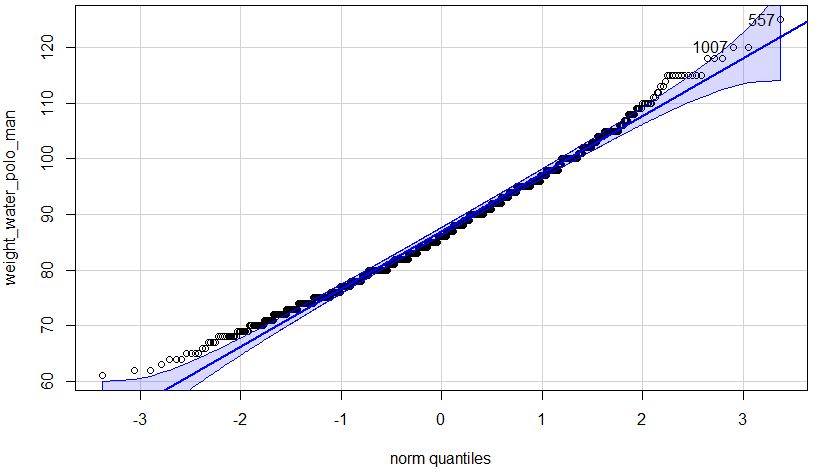
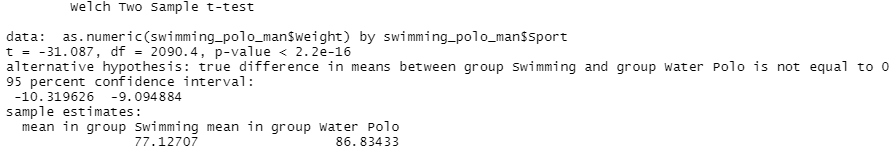


Рис. 6 – Проверка на нормальность веса.

Проведем тест на равенство дисперсий для для пловцов и спортсменов водного поло веса мужчин. Гипотеза H0 – дисперсии равны – принимается за основную:



Проверим вопрос: различаются ли эти средние значения статистически, на основе гипотезы об отсутствии разницы при помощи *t*-теста:



Можем сделать вывод о наличии существенной разницы между средними. Однако такое совпадение выводов будет иметь место не всегда и, следовательно, на разницу между групповыми дисперсиями (или ее отсутствие) следует обращать серьезное внимание при выборе и интерпретации того или иного варианта t-теста.

Вывод: Ознакомился с некоторыми статистическими тестами, принципами их работы. Научился оценивать нормальность распределения выборки, а также выполнил оценку статистических гипотез.