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 №7**

**Дисциплина: Обработка больших данных**

Работу выполнила: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Д. Н. Баева

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В. И. Шиян

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

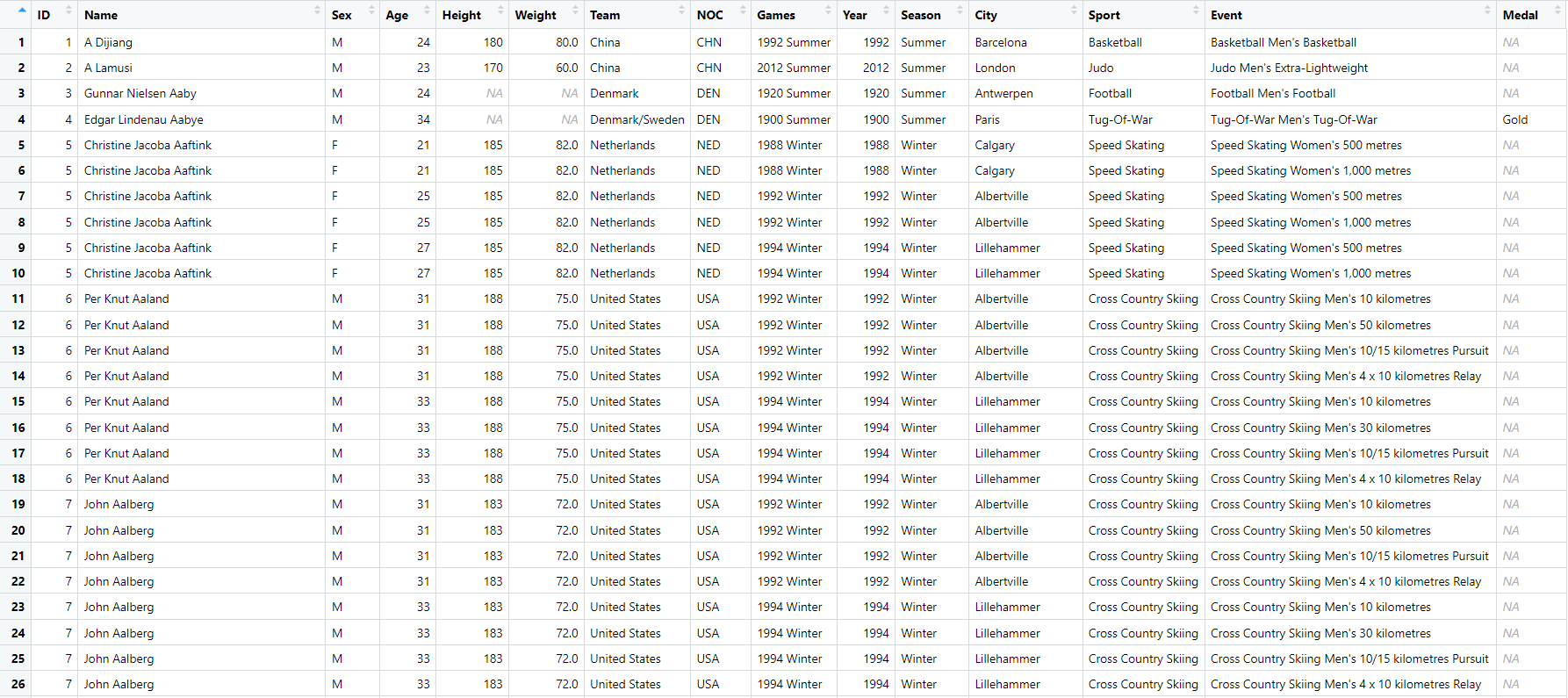
Из исходного csv-файла были импортированы данные в RStudio. Таблица представлена на рисунке 1.

Рисунок 1 – Исходная таблица с данными

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, а также повторяющиеся строчки, и строчки с пустым значением поля “Вес”. На рисунке 2 изображена таблица получившихся данных.

Рисунок 2 – Измененная таблица с данными

Затем 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.

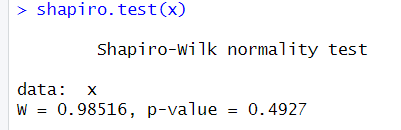


Рисунок 3 – Результат теста Шапиро-Уилкса

Значение p-value равно 0.4927. Так как это значение больше, чем граничное значение 0.05, то гипотеза не отвергается, а значит распределение нормальное. Визуальное подтверждение изображено на рисунках 4,5.

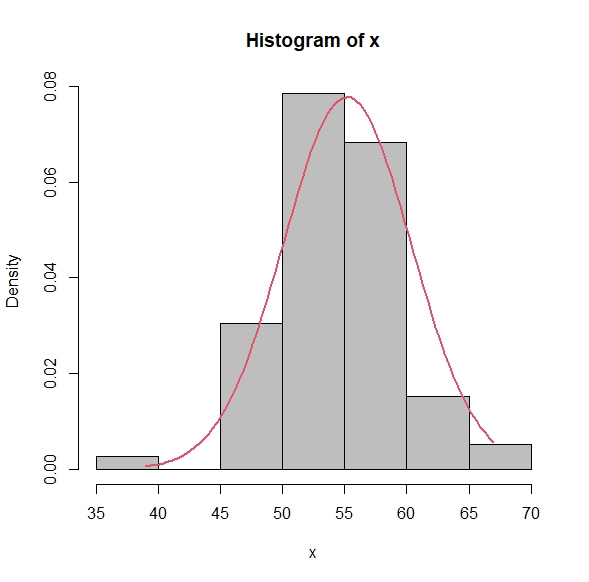


Рисунок 4 – Гистограмма с линией плотности

Изображение выглядит как текст, линия, диаграмма, График

Автоматически созданное описание

Рисунок 5 – Квантильно-квантильный график

Далее проверим гипотезу о среднем весе спортсменок, занимающихся синхронным плаванием. Первая гипотеза гласит: “ Средний вес спортсменок, занимающихся синхронным плаванием, составляет 50 килограмм. На рисунке 6 показан результат выполнения теста.

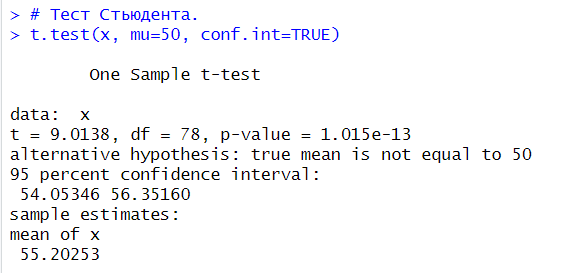


Рисунок 6 – Результат выполнения теста Стьюдента

Значение p-value <0.05, поэтому гипотеза отвергается. Доверительный интервал расположен от 54.05 до 56.35, именно в этот интервал входит настоящий средний вес спортсменок – 55.20.

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

Сперва вновь данные проверяются на нормальность распределения тестом Шапиро-Уилкса. На рисунке 7 представлен результат этого теста.

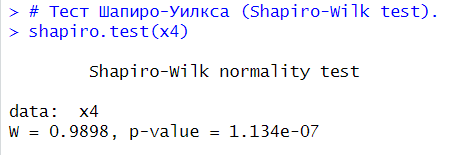


Рисунок 7 – Результат выполнения теста Шапиро-Уилкса

По результатам теста, значение p-value вновь меньше граничного 0.05. Следовательно, гипотеза о нормальном распределении не подтвердилась. Для большей наглядности построим графики для данных по синхронному плаванию и плаванию. График изображён на рисунке 8.

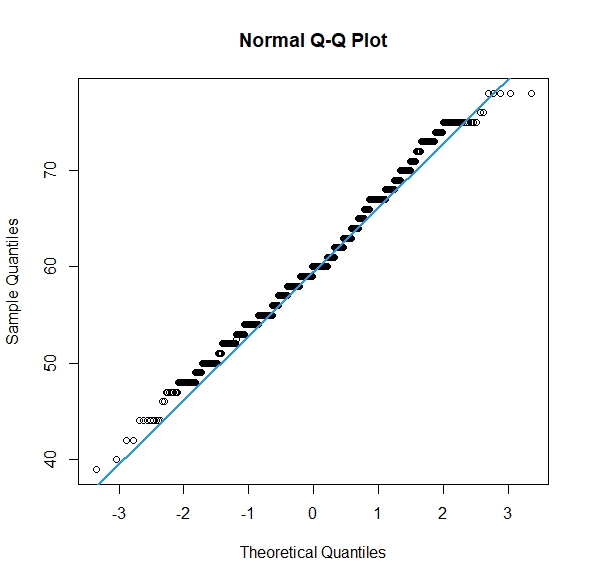


Рисунок 8 – Квантильно-квантильный график для двух видов спорта

График подтверждает, что данные не распределены нормально.

Затем, непосредственно проведем применим тест Флингера-Киллина для проверки равенства веса. Он непараметрический и отлично подходит для анализа данных с ненормальным распределением. На рисунке 9 изображён результат теста.

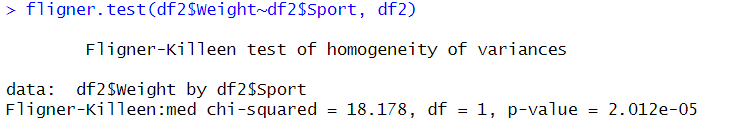


Рисунок 9 – Результаты выполнения теста Флингера-Киллина

Видно, что значение p-value близко к нулю, поэтому отвергаем нулевую гипотезу о равенстве веса. Для того, чтобы узнать приблизительные значения среднего веса спортсменок для двух видов спорта, применим двух выборочный тест Стьюдента. На рисунке 10 показан результат теста.

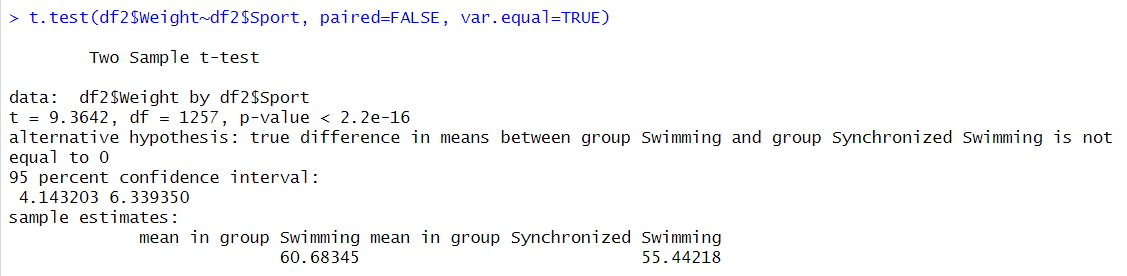


Рисунок 10 – Результат работы t-test

Явно видно, что значение p-value близко к нулю, а значит гипотеза не верна и равенство среднего веса спортсменок синхронного плавания и плавания не подтверждено. Этот тест выводит также средний вес, поэтому можно наглядно убедиться, что средний вес спортсменок по синхронному плаванию меньше на 5 килограммов веса спортсменок по плаванию.

**Вывод:** В данной работе были ознакомлены с некоторыми статистическими тестами, принципами их работы. Помимо этого, было изучено оценивание нормальности распределения выборки, а также выполнение оценки статистических гипотез.

**ПРИЛОЖЕНИЕ А**

**Листинг программы**

# Задание 1.

# импорт данных (все спортсмены по всем видам спорта)

df <- read.table("C:/Users/Дианочка/Desktop/Обработка больших данных/7LW/athlete\_events.csv", sep=",", header=TRUE)

View(df)

v<-c("Name", "Sex", "Weight", "Sport")

df1<-df[which(df[, "Sport"]=="Synchronized Swimming"), v]

# удаление строчек с пустым значением поля "Вес"

df1<-df1[-which(is.na(df1[, "Weight"])),]

# удаление повторяющихся строчек

df1<-unique(df1)

# данные после нормализации

View(df1)

# значения столбца "Вес" становятся вектором

x<-df1[1:nrow(df1), "Weight"]

# Одномерные статистические тесты.

# проверка на нормальность распределения

# Тест Шапиро-Уилкса (Shapiro-Wilk test).

shapiro.test(x)

# Графический способ.

# гистограмма с линией плотности

x2<-seq(min(x), max(x), length=length(x))

fun<-dnorm(x2, mean=mean(x), sd=sd(x))

hist(x, freq=FALSE, col="gray")

lines(x2, fun, col=2, lwd=2)

# квантильно-квантильный график

qqnorm(x)

qqline(x, col=4, lwd=2)

# Тест Стьюдента.

t.test(x, mu=50, conf.int=TRUE)

# Тест Уилкоксона.

wilcox.test(x, mu=mean(x), conf.int=TRUE)

#Задание 2.

df2<-df[which(df[, "Sport"]%in%c("Synchronized Swimming", "Swimming")), v]

df2<-df2[-which(is.na(df2[, "Weight"])),]

df2<-df2[which(df2[, "Sex"]=="F"),]

df2$Sport <- factor(df2$Sport)

df2$Sport <- droplevels(df2$Sport)

x4<-df2[1:nrow(df2), "Weight"]

# проверка на нормальность распределения

# Тест Шапиро-Уилкса (Shapiro-Wilk test).

shapiro.test(x4)

# квантильно-квантильный график

qqnorm(x4)

qqline(x4, col=4, lwd=2)

# проверка равенство дисперсий

# Тест Флингера-Киллина.

fligner.test(df2$Weight~df2$Sport, df2)

# проверка на отсутствие разницы среднестатистическом значении

t.test(df2$Weight~df2$Sport)

t.test(df2$Weight~df2$Sport, paired=FALSE, var.equal=TRUE)