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»

Выполнил: ст. гр. 39/1

Кочнев В.Ю.

Проверил: преподаватель

Яхонтов А.А.

Краснодар

2025

Тема: Проверка статистических гипотез

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

Задание:

1. Провести дескриптивный анализ данных.

2. Провести проверку на нормальность и дисперсию. Из чего сделать вывод о требуемом тесте.

3. Проверьте гипотезу о среднем весе спортсменов выбранного вида спорта (вид спорта остается по ЛР4).

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

Ход работы:

1) Дан датасет, необходимо провести дескриптивный анализ данных

data <- read.csv("athlete\_events.csv")

dataset <- data[data$Sport == "Sailing",]

datasetAgeWeightHeight <- data.frame(dataset$Age, dataset$Height, dataset$Weight)

datasetMan <- dataset[dataset$Sex == "M", ]

datasetFemale <- dataset[dataset$Sex == "F", ]

datasetManAgeWeightHeight <- data.frame(datasetMan$Age, datasetMan$Height, datasetMan$Weight)

datasetFemaleAgeWeightHeight <- data.frame(datasetFemale$Age, datasetFemale$Height, datasetFemale$Weight)

# Провести дескриптивный анализ

min <- apply(dataset, 2, min, na.rm = TRUE)

max <- apply(dataset, 2, max, na.rm = TRUE)

mean <- apply(datasetAgeWeightHeight, 2, mean, na.rm = TRUE)

minMan <- apply(datasetMan, 2, min, na.rm = TRUE)

maxMan <- apply(datasetMan, 2, max, na.rm = TRUE)

meanMan <- apply(datasetManAgeWeightHeight, 2, mean, na.rm = TRUE)

minFeamle <- apply(datasetFemale, 2, min, na.rm = TRUE)

maxFemale <- apply(datasetFemale, 2, max, na.rm = TRUE)

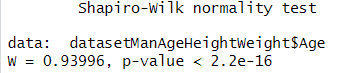
meanFemale <- apply(datasetFemaleAgeWeightHeight, 2, mean, na.rm = TRUE)

2) Провести проверку на нормальность и дисперсию.

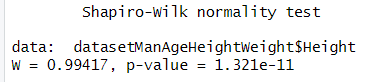
Проверяться будут возраст, рост и вес, по женщинам и мужчинам(Числовые параметры) по парусному спорту

Тест шапиро

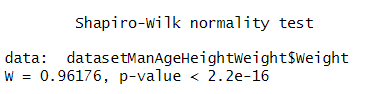
shapiro.test(datasetManAgeHeightWeight$Age)



shapiro.test(datasetManAgeHeightWeight$Height)

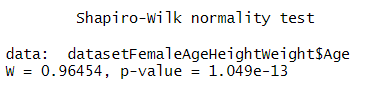


shapiro.test(datasetManAgeHeightWeight$Weight)

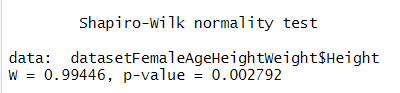


По всем полям не нормальное распределение так вероятность ошибки 1-го рода близка к нулю

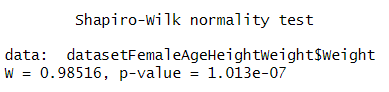
shapiro.test(datasetFemaleAgeHeightWeight$Age)



shapiro.test(datasetFemaleAgeHeightWeight$Height)



shapiro.test(datasetFemaleAgeHeightWeight$Weight)

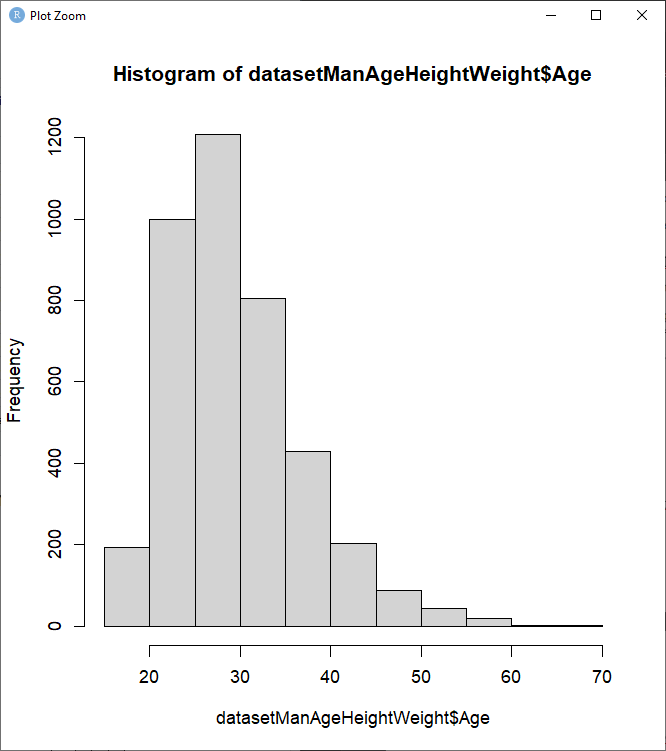


Данные так же не нормальны

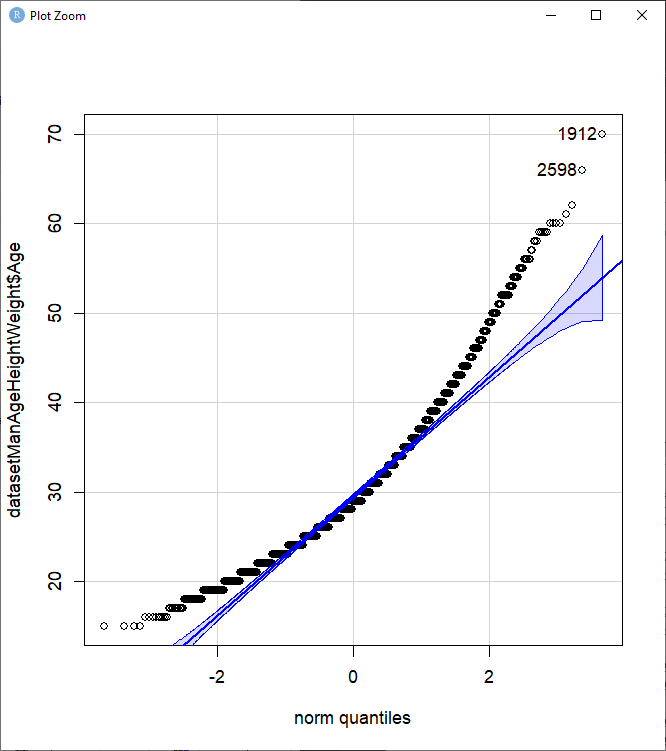
Графическая проверка

library('car')

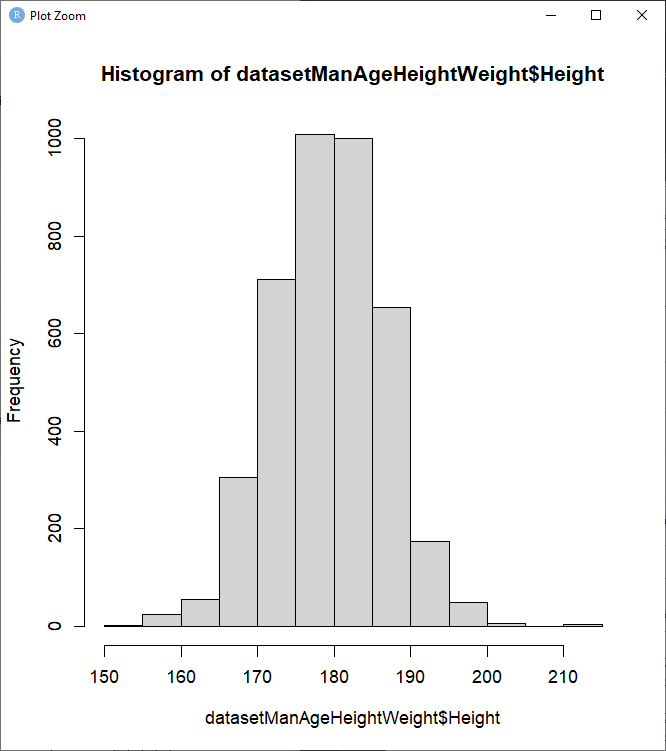
hist(datasetManAgeHeightWeight$Age)



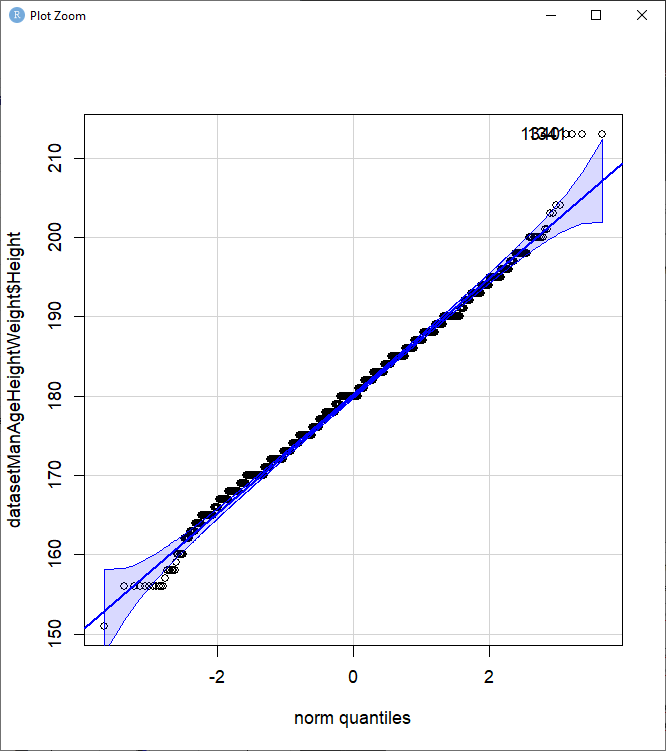
qqPlot(datasetManAgeHeightWeight$Age)



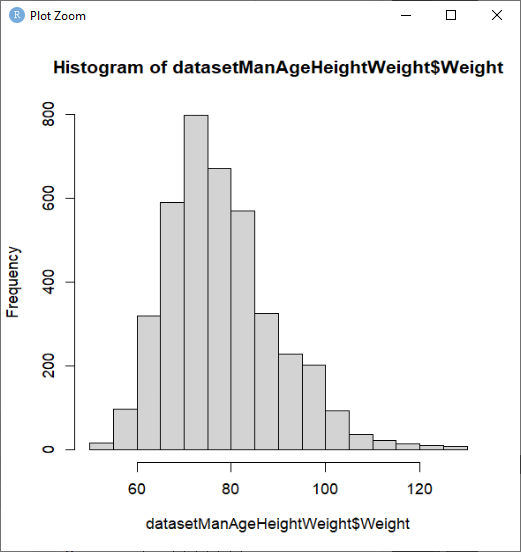
hist(datasetManAgeHeightWeight$Height)



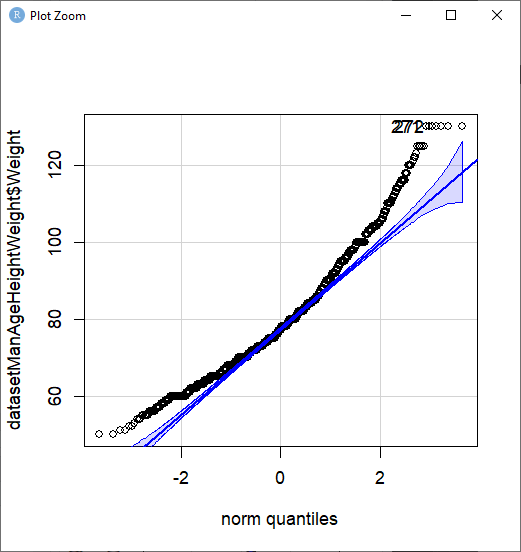
qqPlot(datasetManAgeHeightWeight$Height)



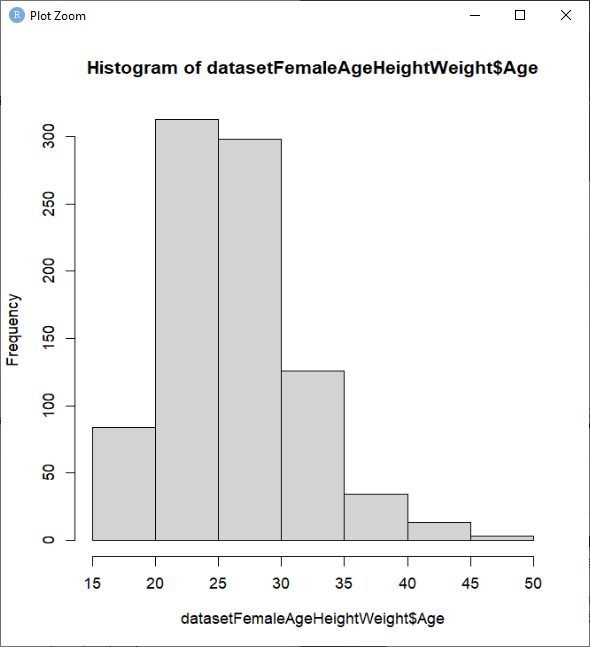
hist(datasetManAgeHeightWeight$Weight)



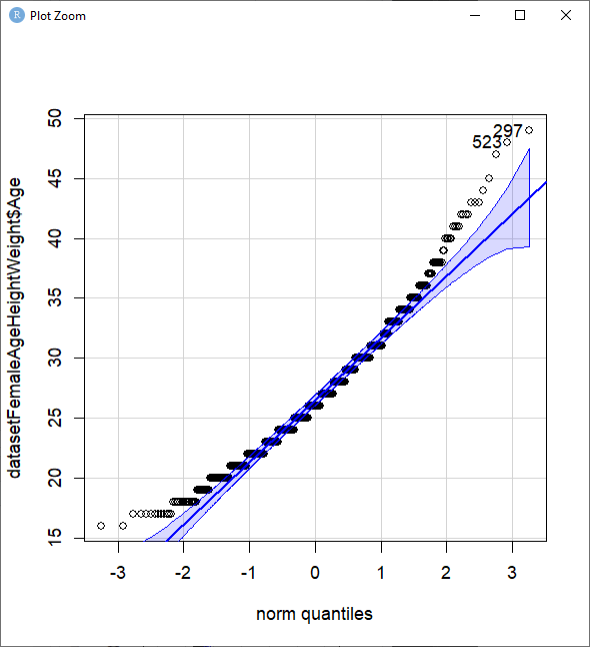
qqPlot(datasetManAgeHeightWeight$Weight)



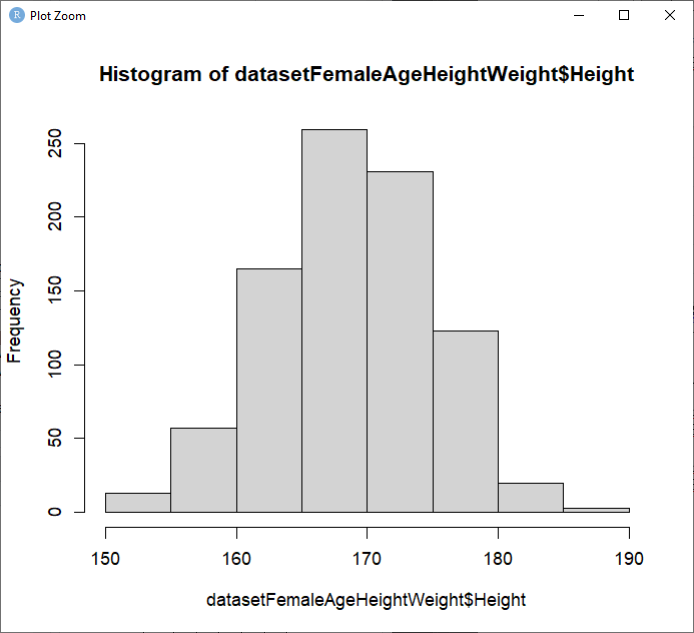
hist(datasetFemaleAgeHeightWeight$Age)



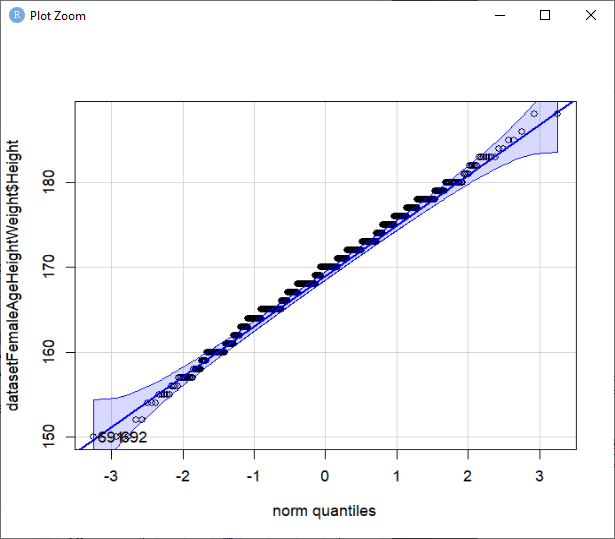
qqPlot(datasetFemaleAgeHeightWeight$Age)



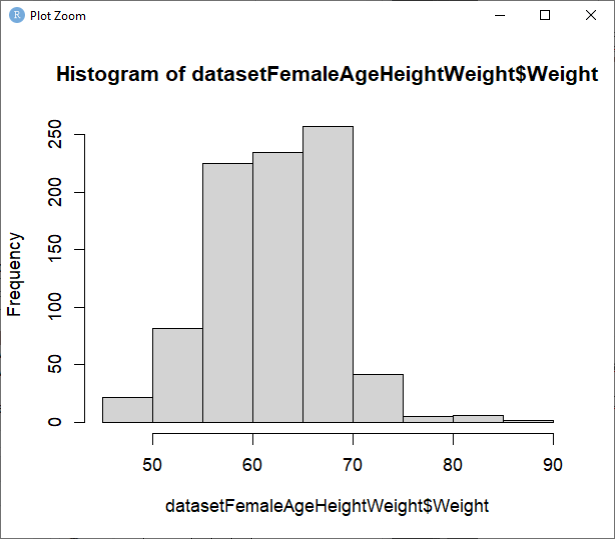
hist(datasetFemaleAgeHeightWeight$Height)



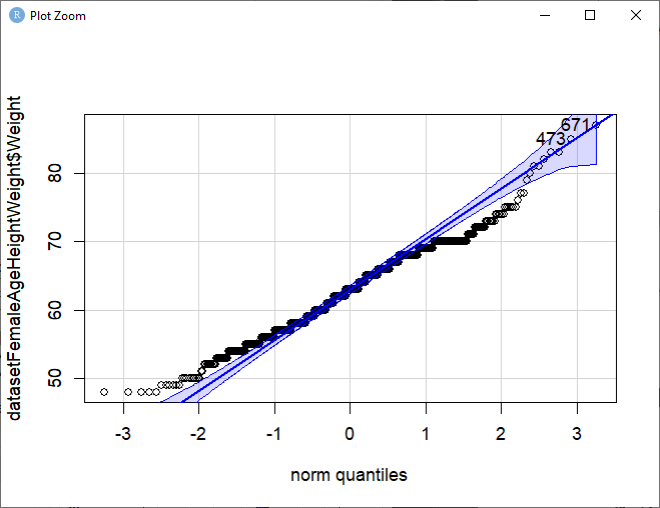
qqPlot(datasetFemaleAgeHeightWeight$Height)



hist(datasetFemaleAgeHeightWeight$Weight)



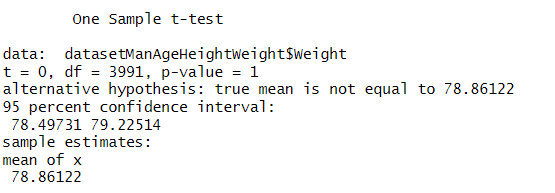
qqPlot(datasetFemaleAgeHeightWeight$Weight)



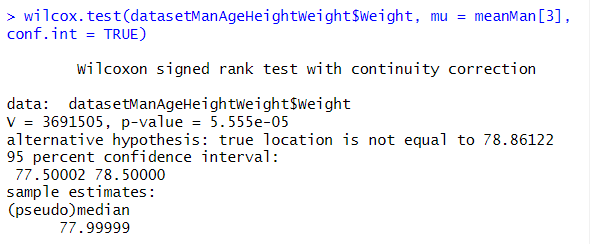
4) Проверьте гипотезу о среднем весе спортсменов парусного спорта

Студент тест для мужчин и женщин

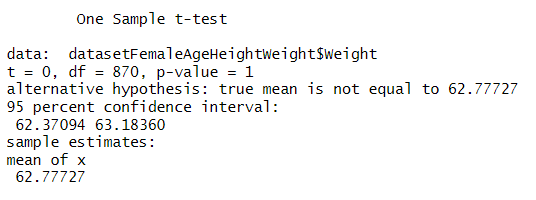
t.test(datasetManAgeHeightWeight$Weight, mu = meanMan[3])



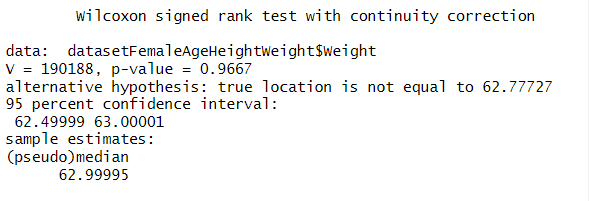
wilcox.test(datasetManAgeHeightWeight$Weight, mu = meanMan[3], conf.int = TRUE)



t.test(datasetFemaleAgeHeightWeight$Weight, mu = meanFemale[3])



wilcox.test(datasetFemaleAgeHeightWeight$Weight, mu = meanFemale[3], conf.int = TRUE)



5) Проверьте гипотезу о равенстве среднего веса женщин (мужчин) в двух разных выбранных видах спорта

dataset2 <- data[data$Sport == "Basketball",]

dataset2SportAgeHeightWeight <- data.frame( dataset2$Sport,dataset2$Age, dataset2$Height, dataset2$Weight)

dataset2SportAgeHeightWeight <- dataset2SportAgeHeightWeight[rowSums(is.na(dataset2SportAgeHeightWeight)) == 0,]

colnames(dataset2SportAgeHeightWeight) = c("Sport","Age", "Height", "Weight")

dataset2AgeHeightWeight<-dataset2SportAgeHeightWeight[2:length(dataset2SportAgeHeightWeight)]

dataset2Man <- dataset2[dataset2$Sex == "M", ]

dataset2Female <- dataset2[dataset2$Sex == "F", ]

dataset2SportManAgeHeightWeight <- data.frame(dataset2Man$Sport,dataset2Man$Age, dataset2Man$Height, dataset2Man$Weight)

dataset2SportManAgeHeightWeight <- dataset2SportManAgeHeightWeight[rowSums(is.na(dataset2SportManAgeHeightWeight)) == 0,]

colnames(dataset2SportManAgeHeightWeight) = c("Sport","Age", "Height", "Weight")

dataset2ManAgeHeightWeight<-dataset2SportManAgeHeightWeight[2:length(dataset2SportManAgeHeightWeight)]

dataset2SportFemaleAgeHeightWeight <- data.frame(dataset2Female$Sport,dataset2Female$Age, dataset2Female$Height, dataset2Female$Weight)

dataset2SportFemaleAgeHeightWeight <- dataset2SportFemaleAgeHeightWeight[rowSums(is.na(dataset2SportFemaleAgeHeightWeight)) == 0,]

colnames(dataset2SportFemaleAgeHeightWeight) = c("Sport","Age", "Height", "Weight")

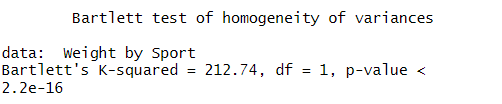
dataset2FemaleAgeHeightWeight<-dataset2SportFemaleAgeHeightWeight[2:length(dataset2SportFemaleAgeHeightWeight)]

# Проверка на нормальность, дисперсию и равсентсве среднего в двух разных видах спорта

combdatasetFemale <- rbind(datasetSportFemaleAgeHeightWeight, dataset2SportFemaleAgeHeightWeight)

hist(combdatasetFemale$Weight)

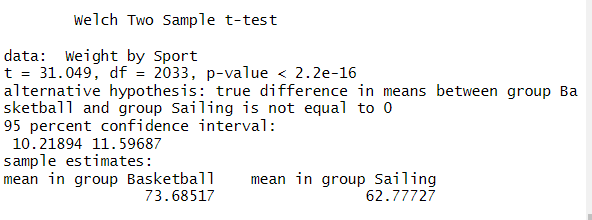
bartlett.test(Weight~Sport, data = combdatasetFemale)



Дисперсии различны

Проверка равенства средниего при уловии что дисперсии разные

t.test(Weight~Sport, data = combdatasetFemale)



Средние разные и находятся в диапозоне от 10.2 до 11.6

Проверка равенства средниего при уловии что дисперсии равные хотя это не так

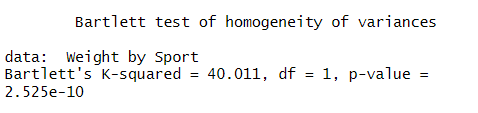
t.test(Weight~Sport, data = combdatasetFemale, ver.equal = TRUE)

Проверка мужчин

combdatasetMan <- rbind(datasetSportManAgeHeightWeight, dataset2SportManAgeHeightWeight)

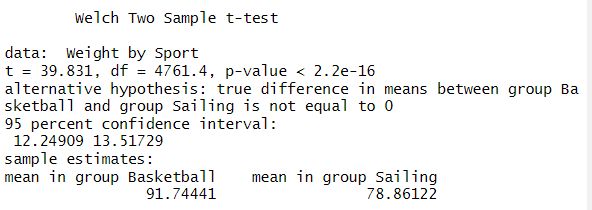
hist(combdatasetMan$Weight)

bartlett.test(Weight~Sport, data = combdatasetMan)



Дисперсии разные

t.test(Weight~Sport, data = combdatasetMan)



Средние разные и находятся на растоянии в диапозоне от 12.4 до 13.5

t.test(Weight~Sport, data = combdatasetMan, ver.equal = TRUE)

Вывод: В ходе лабораторной работы я ознакомился с некоторыми статистическими тестами, принципами их работы. Научился оценивать нормальность распределения выборки, а также выполнил оценку статистических гипотез.