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 выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ К. А. Корнилов

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_В. И. Шиян

**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. Был выполнен дискриптивный анализ набора данных об олимпиадах и их участниках.

Код:

data = data.frame(read.csv('lab7/athlete\_events.csv'))

summary(data)

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

mean(data$Height[!is.na(data$Height)])

data$Height[is.na(data$Height)]=mean(data$Height[!is.na(data$Height)])

data$Weight[is.na(data$Weight)]=mean(data$Weight[!is.na(data$Weight)])

data$Age[is.na(data$Age)]=mean(data$Age[!is.na(data$Age)])

data$Medal[is.na(data$Medal)]="None"

modas = apply(data,2,getmode)

modas

variances = apply(data,2,var)

variances

sds = apply(data,2,sd)

sds

#Делаем графики боксплотов

par(mfrow=c(2,2))

boxplot(data$Weight,main="weight")

boxplot(data$Height,main="Height")

boxplot(data$Age,main="age",min=10,max=97)

#Смотрим распределение

par(mfrow=c(4,4))

for(i in colnames(data)){

if(is.numeric(data[,i])){

hist(data[,i],main=i)

}

else{

f=factor(data[,i])

sort(levels(f))

levels(f) = sort(levels(f))

as.numeric(f)

hist(as.numeric(f),main=i,xaxt="n")

axis(1,at=1:length(levels(f)),labels=unique(levels(f)))

}

}

Результаты выполнения дескриптивного анализа представлена на рисунках 1 – 7. Согласно данным исследованиям видно,что вес, возраст и рост спортсменов приближаются к смещенному нормальному распределению. Также видно, что женщин – участниц намного меньше участников – мужчин.

Еще видно, что нет спортсмена, участвовавшего в олимпиаде большое количество раз. Заметна также диспропорция в плане городов – олимпиадников. Некоторые города принимали у себя олимпиаду намного больше раз, чем остальные. Также можем заметить, что возраст спортсменов относительно сгруппирован вокруг среднего значения с максимальным отклонением на 6 лет. Вес и рост же спортсменов более размыт со средним отклонением в 10 и 12 соответственно.

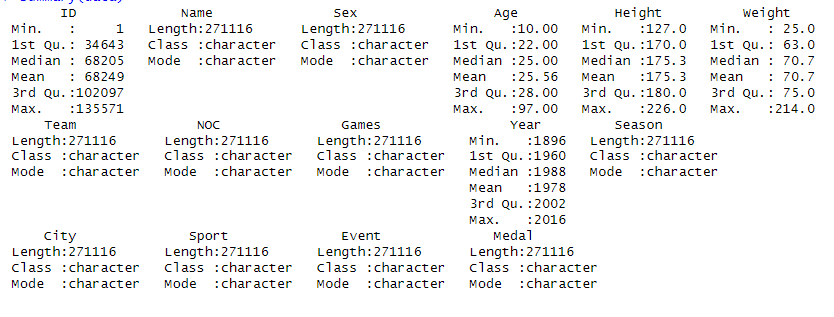


Рисунок 1 – Результаты измерения квартилей для набора данных

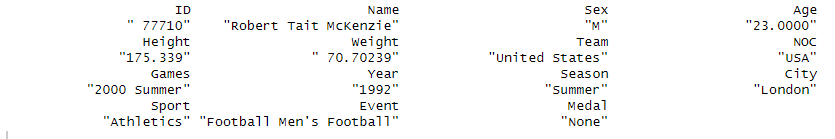


Рисунок 2 – Результаты измерения моды для набора данных

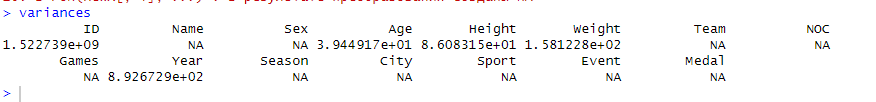


Рисунок 3 – Результаты измерения дисперсии

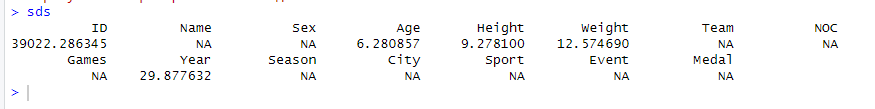


Рисунок 4 – Результаты измерения среднего квадратичного отклонения

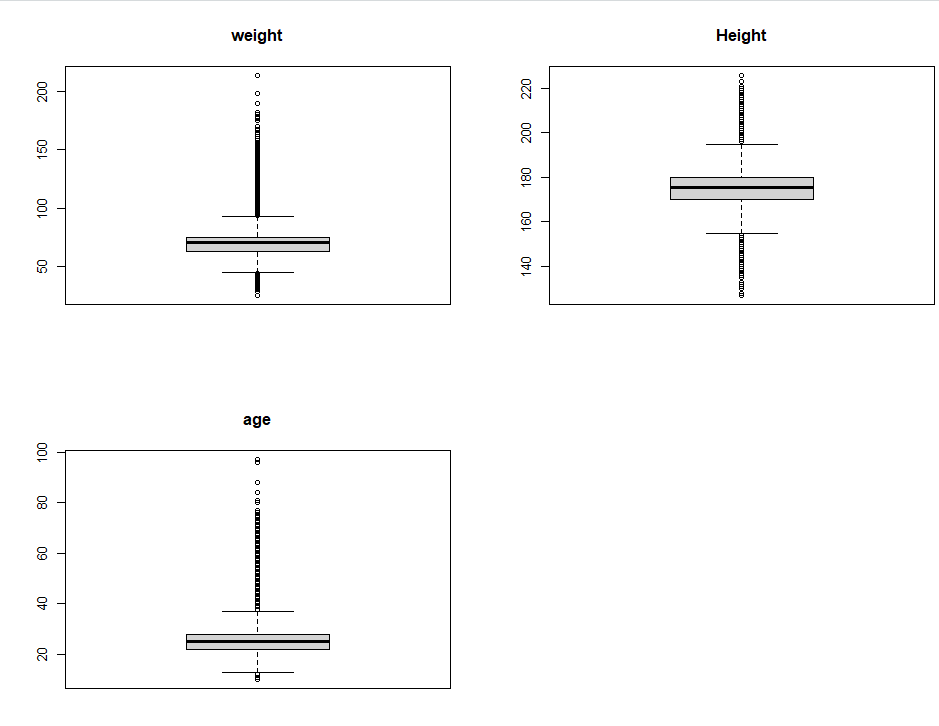


Рисунок 5 – боксплоты с замененными пропущенными данными

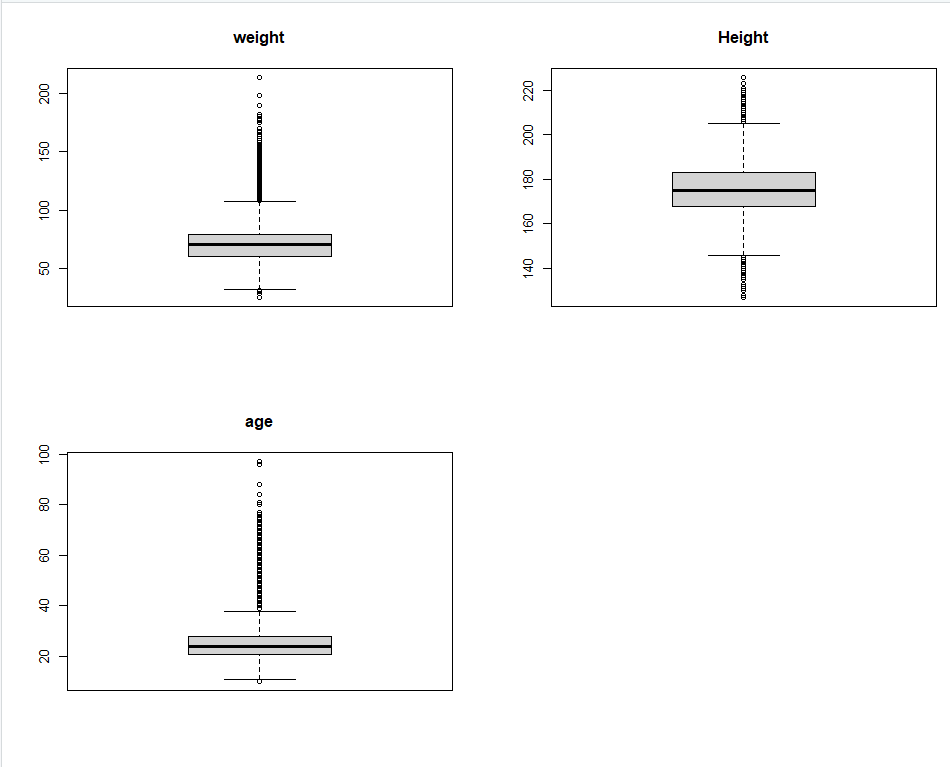


Рисунок 6 – боксплоты без учета пропущенных данных

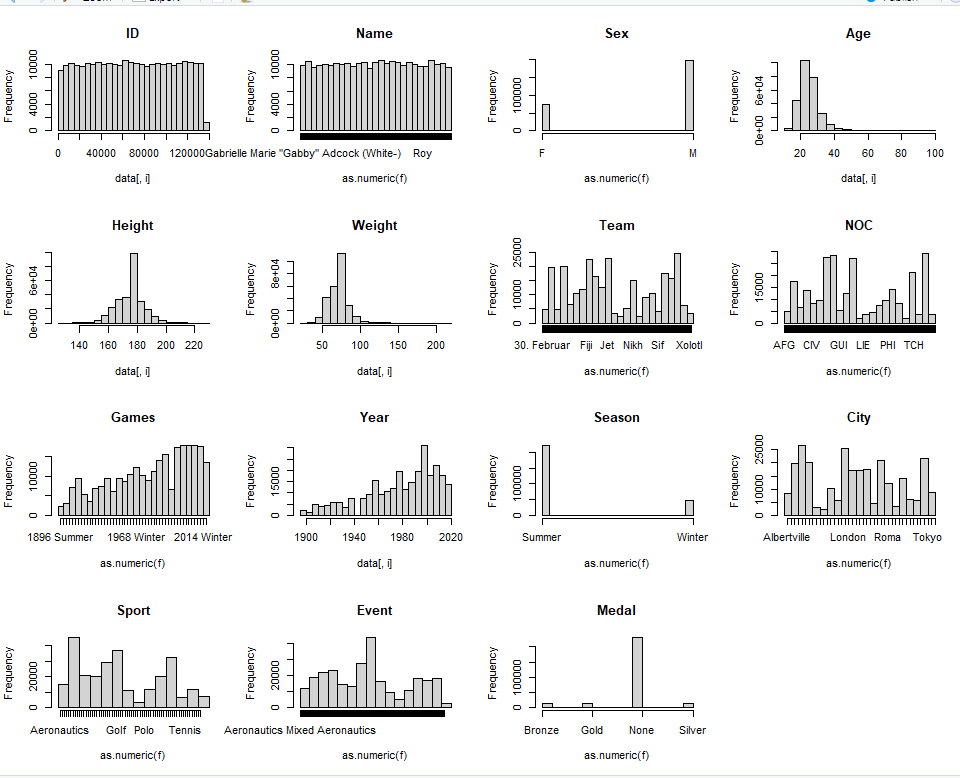
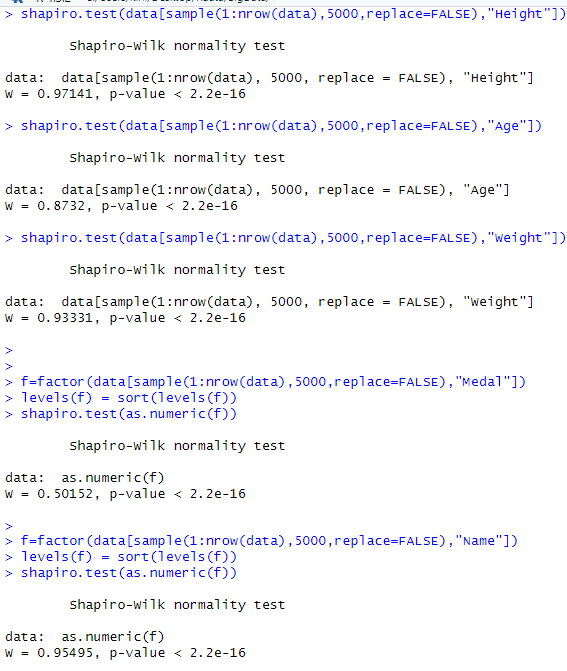
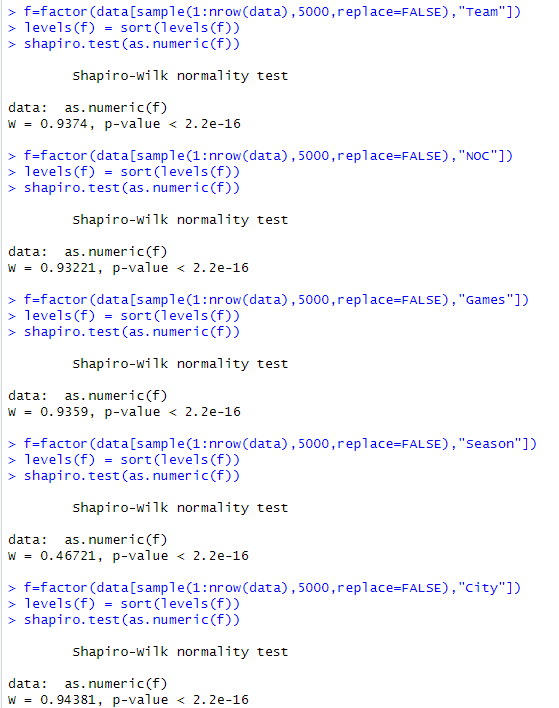
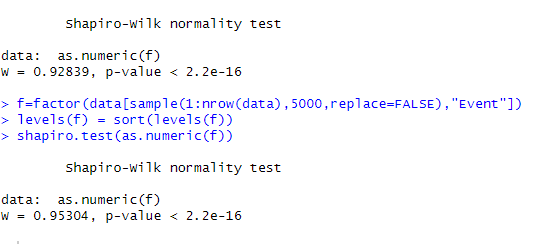


Рисунок 7 – Распределение всех столбцов набора данных

1. Была проведена проверка на нормальность распределения данных c помощью теста Шапиро и графическим методом. Результаты представлены на рисунке 8-10 и 11. Как видно из полученных результатов, ни одна из величин не является нормально распределенной. Но согласно графическому методу вес спортсменов наиболее близко приближается к нормальному распределению.







Рисунки 8-10 – Результаты тестирования на нормальность с помощью теста Шапиро

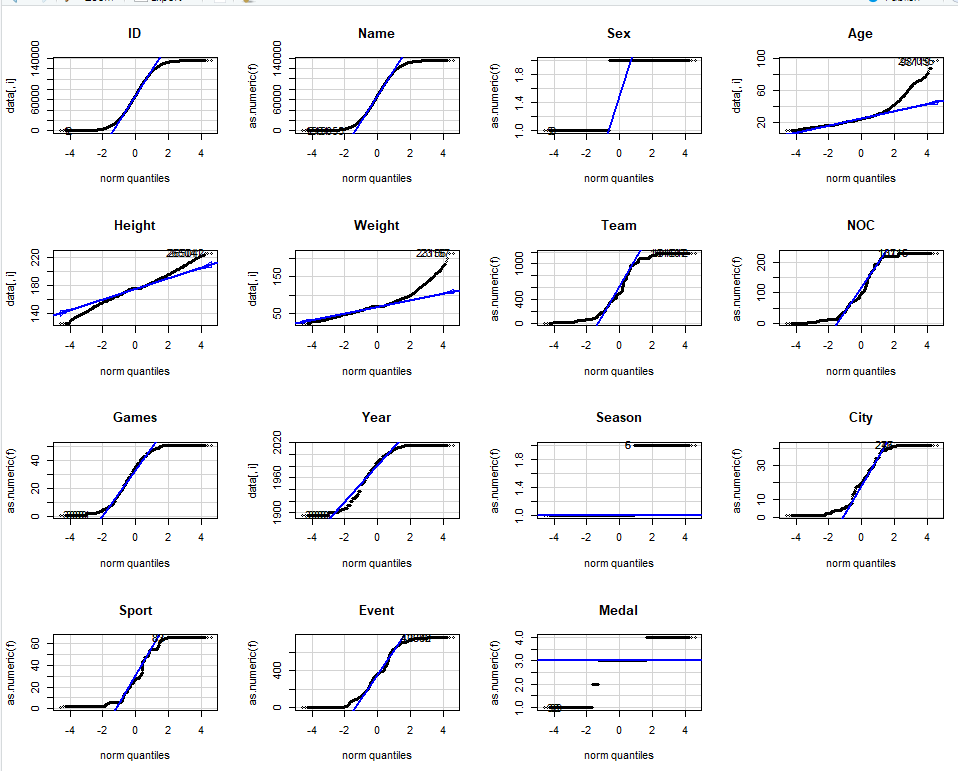


Рисунок 11 – Результаты тестирования графическим методом

1. Была проверена гипотеза на среднее значение веса спортсменов – гимнастов.

Код:

#Проверяем средний вес

data\_gym=data[data$Sport=="Gymnastics","Weight"]

data\_gym

t.test(data\_gym,mu=55)

#Проведем тест без учета,что величина нормальная

wilcox.test(data\_gym,mu=61.999,conf.int=TRUE)

Результаты тестирования представлена на рисунке 12. Поскольку изначальная величина веса согласно графическому теста сильно похожа на нормальное распределение, то проведем для нашей подвыборки тест Стьюдента, предположив, что и подвыборка веса имеет нормальное распределение. Согласно результату p-value < 5%. Следовательно отметаем гипотезу, что наша величина имеет среднее значение 61.1 при допущении о нормальном распределении. Проведем теста Вилксона, предположив,что величина не имеет нормального распределения. Согласно результату p-value>5%. Следовательно, будем придерживаться гипотезы, что величина веса имеет среднее значение в 62.

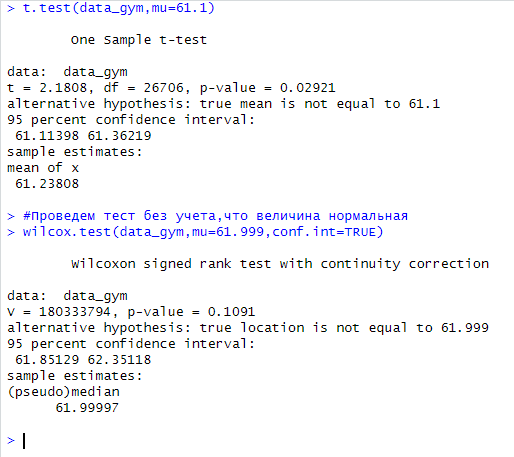


Рисунок 12 – Результаты тестирования гипотезы о среднем значении

1. Проведем двухвыборочный тест на сравнения средних значений женщин – спортсменок двух видов спорта, и аналогично для мужчин.

Код:

data\_box\_f=data[data$Sport=="Boxing"&data$Sex=="F","Weight"]

data\_box\_m=data[data$Sport=="Boxing"&data$Sex=="M","Weight"]

data\_gym\_f=data[data$Sport=="Gymnastics"&data$Sex=="F","Weight"]

data\_gym\_m=data[data$Sport=="Gymnastics"&data$Sex=="M","Weight"]

var(data\_gym\_f)

var(data\_box\_f)

var(data\_gym\_m)

var(data\_box\_m)

shapiro.test(data\_gym[sample(1:nrow(data),5000,replace=FALSE)])

shapiro.test(data\_box[sample(1:nrow(data),5000,replace=FALSE)])

t.test(data\_gym\_f,data\_box\_f)

t.test(data\_gym\_m,data\_box\_m)

wilcox.test(data\_gym\_f,data\_box\_f)

wilcox.test(data\_gym\_m,data\_box\_m)

Результаты выполнения теста представлены на рисунке 13-15. Согласно рисунку 13 сравниваемые выборки имеют разные дисперсии. Поэтому проведем модифицированные версии тестов, предполагающих разные дисперсии у выборок. Согласно рисунку 14 ни одна из выборок не является нормальной, поэтому при проведении теста Стьюдента будем иметь это в виду. Согласно рисунку 15 тест Стьюдента, как и ожидалось, выдал результаты для отбрасывания гипотезы. Однако тест Уильсона для выборки мужчин выдал, что гипотеза ошибочна с p>0.5. А значит, мы можем придерживаться данной гипотезы.

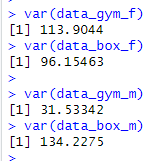


Рисунок 13 – Значения дисперсий полученных выборок

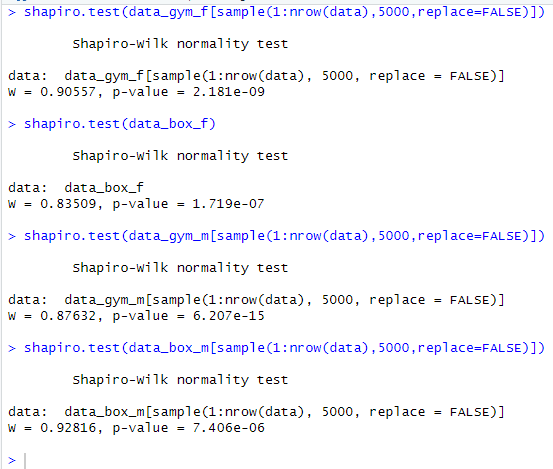


Рисунок 14 – Результаты проверки на нормальность

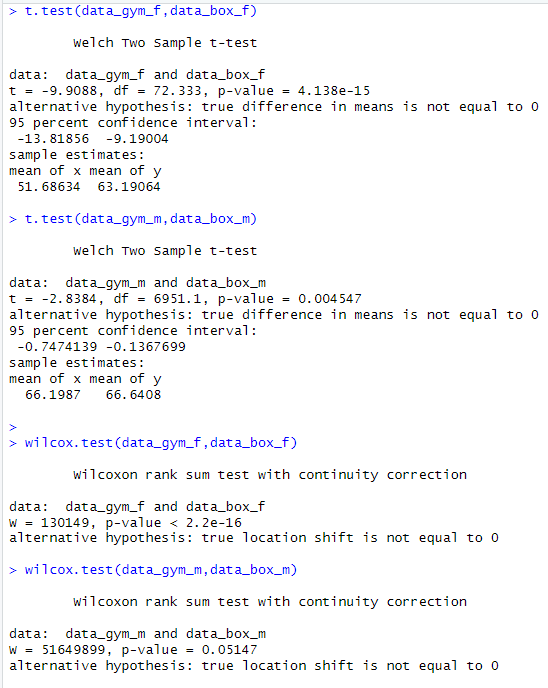


Рисунок 15 – Результаты тестирования

**Вывод:** Проведено ознакомление с некоторыми статистическими тестами, принципами их работы. Получены навыки оценивания нормальности распределения выборки, а также выполнения оценки статистических гипотез.