# Лабораторный практикум ВКИАД c R

**Преподаватель**

Дорошко Ольга Валерьевна

**Студент**

Лозюк Илья Дмитриевич, 2 курс, 7 группа, ПМ

**Вариант № 2**

# Задание 1. Управление данными в R

Загрузка данных из текстового файла в переменную - таблицу данных

qc=read.table("QC2.txt")

Размерность таблицы данных

dim(qc)

## [1] 100 9

Первые три записи из таблицы

head(qc, 3)

## V1 V2 V3 V4 V5 V6 V7 V8 V9  
## 1 56.08 4 2 1 4 31.79 18.417 5.290 2  
## 2 57.00 1 3 1 2 31.20 19.140 4.500 4  
## 3 56.93 5 1 2 4 34.22 18.477 5.285 3

Таблица после переименования переменных

names(qc) = c("c1", "kit", "nbug", "maker", "vendor", "c2", "c3", "c4", "bug")  
head(qc, 3)

## c1 kit nbug maker vendor c2 c3 c4 bug  
## 1 56.08 4 2 1 4 31.79 18.417 5.290 2  
## 2 57.00 1 3 1 2 31.20 19.140 4.500 4  
## 3 56.93 5 1 2 4 34.22 18.477 5.285 3

Комплексная статистика по переменной vendor

summary(qc$vendor)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 1.75 2.00 2.49 4.00 4.00

Комплексная статистика по переменной vendor после ее преобразования в переменную-фактор

summary(as.factor(qc[,"vendor"]))

## 1 2 3 4   
## 25 28 20 27

Число изделий, у которых число некритических дефектов (nbug) больше 3

sum(qc$nbug > 3)

## [1] 11

Наблюдения, для которых количество некритических дефектов равно максимальному

qc[max(qc[,"nbug"]) == qc[,"nbug"],]

## c1 kit nbug maker vendor c2 c3 c4 bug  
## 15 53.93 3 6 2 1 31.69 18.006 5.080 3  
## 72 57.19 7 6 1 1 30.27 18.930 4.925 3

# Задание 2. Описательная статистика и графический анализ

Присоединение таблицы данных к списку текущих переменных и загрузка дополнительной библиотеки

attach(qc)  
library("e1071")

Выборочные среднее и медиана для переменной с1 в указанном порядке

mean(c1)

## [1] 55.0549

median(c1)

## [1] 54.855

Выборочные квартили Q1,Q2,Q3 , а также интерквартильный размах для переменной c1

quantile(c1, seq(0.25, 0.75, 0.25))

## 25% 50% 75%   
## 53.6525 54.8550 56.5250

IQR(c1)

## [1] 2.8725

Комплексная статистика по переменной c1

summary(c1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 49.68 53.65 54.85 55.05 56.52 59.94

Дисперсия для переменной с1 и переменной, которая представляет собой линейную комбинацию первой

var(c1)

## [1] 4.628207

var(c1 \* 2 + 100)

## [1] 18.51283

**Вывод:** вторая дисперсия больше в 4 раза.

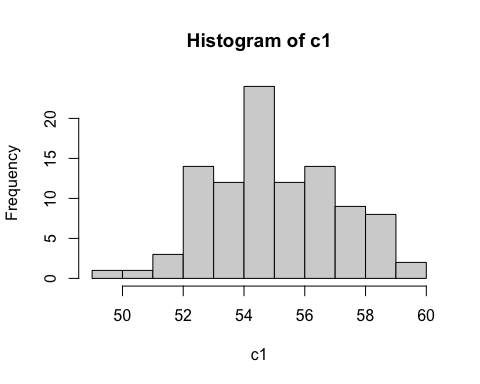
Дисперсия для переменно с1, вычесленная с использованием векторного выражения

sum((mean(c1)-c1)^2/(length(c1) - 1))

## [1] 4.628207

Гистограмма для переменной с1

hist(c1)

 **Вывод:** распределение переменной c1 скошено влево, если сравнивать с нормальным распределением.

Асимметрия и эксцесс переменной c1 в указанном порядке

skewness(c1)

## [1] 0.03317255

kurtosis(c1)

## [1] -0.4727926

**Вывод:** распределение переменной c1 скошено влево и график более пологий, чем график нормального распределения.

Проверка распределения переменной c1 на нормальность с помощью статистического теста Лиллиефорса

library(nortest)  
lillie.test(c1)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: c1  
## D = 0.0772, p-value = 0.151

**Вывод:** гипотеза о нормальном распределении не отклоняется на уровне значимости 0.05.

# Задание 3. Анализ статистических зависимостей

Перекодировка переменной kit в переменную с двумя значениями (1 и 2)

kit = ifelse(kit < 3, 1, 2)

Ранговый коэффициент корреляции Спирмена для переменных kit и maker

cor(kit, maker, method="spearman")

## [1] 0.08173077

Таблица сопряженности для переменных kit и maker

table(kit, maker)

## maker  
## kit 1 2  
## 1 27 21  
## 2 25 27

**Вывод:** согласно ранговому коэффициенту корреляции и таблице сопряженности согласованность между переменными практически отсутствует.

Тест на равенство нулю коэффициента корреляции для переменных с1,с3

cor.test(c1, c3)

##   
## Pearson's product-moment correlation  
##   
## data: c1 and c3  
## t = 21.001, df = 98, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.8611169 0.9348641  
## sample estimates:  
## cor   
## 0.9045418

**Вывод:** гипотеза о равенстве нулю коэффициента корреляции между переменными отклоняется на уровне значимости 0.05.

Корреляционная матрица для числовых характеристик с1-c4

cor(qc[,c("c1", "c2", "c3", "c4")])

## c1 c2 c3 c4  
## c1 1.0000000 -0.20389934 0.9045418 0.38674144  
## c2 -0.2038993 1.00000000 -0.1960740 -0.09152175  
## c3 0.9045418 -0.19607400 1.0000000 0.50604092  
## c4 0.3867414 -0.09152175 0.5060409 1.00000000

Матрица корреляции в удобном представлении

symnum(cor(qc[,c("c1", "c2", "c3", "c4")]))

## c1 c2 c3 c4  
## c1 1   
## c2 1   
## c3 \* 1   
## c4 . . 1   
## attr(,"legend")  
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '\*' 0.95 'B' 1

**Вывод:** наибольший коэффициент корреляции между переменными c1 и с3.

Построение регрессионной модели по переменным c1 (зависимая - отклик) и с3 (независимая - фактор) и вывод общей статистики по результатам оценивания.

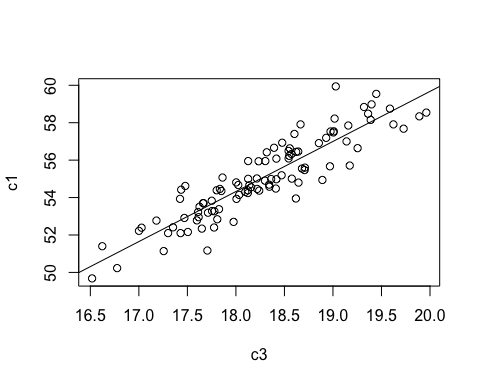
linmod = lm(c1 ~ c3)  
summary(linmod)

##   
## Call:  
## lm(formula = c1 ~ c3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.36553 -0.60136 -0.01133 0.58398 2.86565   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.1747 2.3293 2.651 0.00936 \*\*   
## c3 2.6748 0.1274 21.001 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.922 on 98 degrees of freedom  
## Multiple R-squared: 0.8182, Adjusted R-squared: 0.8163   
## F-statistic: 441 on 1 and 98 DF, p-value: < 2.2e-16

**Вывод:** значим (на уровне 0.05) только коэффициент при независимой переменной c3. Статистика R-квадрат принимает значение 0.8182.

Диаграмма рассеяния с линией регрессии для переменных c1 и c2

plot(c3, c1)  
abline(linmod)



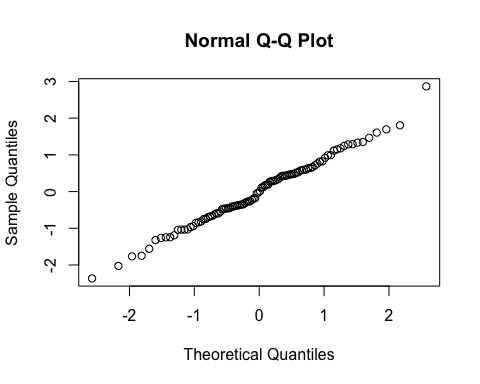
Коэффициенты модели

coefficients(linmod)

## (Intercept) c3   
## 6.174723 2.674845

График “Квантиль-квантиль” для остатков модели

res = residuals(linmod)  
qqnorm(res)

 **Вывод:** большинство точек располагаются вблизи прямой линии, поэтому распределение остатков близко к нормальному.

Тест Колмогорова-Смирнова на нормальность остатков

ks.test(res, pnorm, mean(res), sd(res))

##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: res  
## D = 0.061976, p-value = 0.8371  
## alternative hypothesis: two-sided

**Вывод:** критерий не отклоняет гипотезу о нормальном распределении на уровне значимости 0.05.

# Задание 4. Анализ неоднородных данных

Частоты значений для переменной maker

table(maker)

## maker  
## 1 2   
## 52 48

Двухвыборочный t-критерий по переменной c2 c разбиением на подвыборки по переменной maker

t.test(c2 ~ maker)

##   
## Welch Two Sample t-test  
##   
## data: c2 by maker  
## t = -5.036, df = 95.109, p-value = 2.253e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.942217 -0.843905  
## sample estimates:  
## mean in group 1 mean in group 2   
## 32.02423 33.41729

**Вывод:** гипотеза о равенстве средних в двух подвыборках отклоняется на уровне значимости 0.05 для переменной c2(средние значения характеристики c2 у двух производителей статистически различимы).

“Ящики с усами” для переменной с2 в разрезе по номеру производителя (maker)

boxplot(c2 ~ maker)

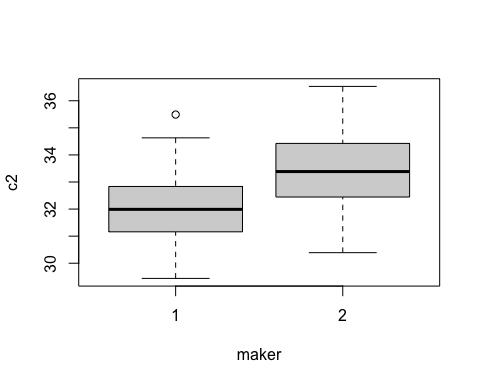
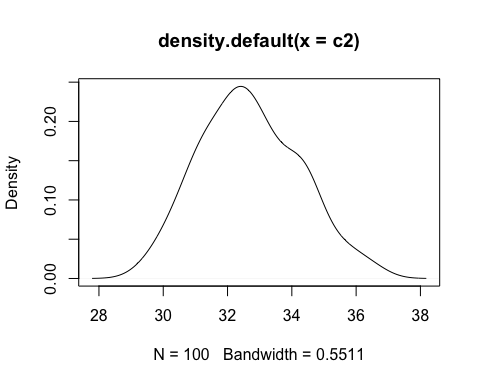


График ядерной оценки плотности распределения для переменной c2

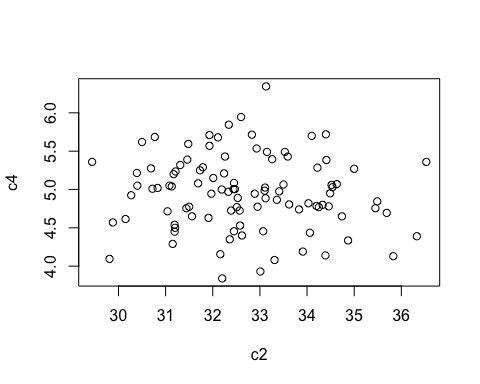
plot(density(c2))

 **Вывод:** из графика видно, что выборка описывается моделью, которая представляет собой смесь двух распределений.

# Задание 5. Классификация неоднородных данных

Диаграмма рассеяния переменных с2 и с4, для которых коэффициент корреляции принимает наиболее близкое к нулю значение

plot(c2, c4)



Первые строки матрицы из переменных c2 и с4

bind1 = cbind(c2, c4)  
head(bind1, 3)

## c2 c4  
## [1,] 31.79 5.290  
## [2,] 31.20 4.500  
## [3,] 34.22 5.285

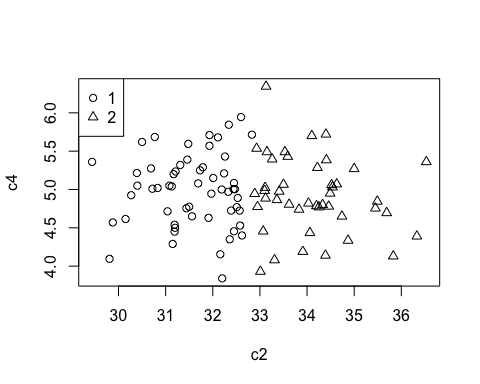
Применение кластерного анализа и вывод частот наблюдений в каждом оцененном классе

kmres = kmeans(bind1, 2)  
table(kmres$cluster)

##   
## 1 2   
## 56 44

График с классовой принадлежностью наблюдений

plot(bind1, pch=kmres$cluster)  
legend("topleft", legend = c(1, 2), pch = c(1, 2))



Работа с данными закончена

detach(qc)