Министерство образования и науки Российской Федерации  
  
Федеральное государственное автономное образовательноеучреждение высшего образования«Самарский государственный аэрокосмический университет имени академика С.П. Королева(национальный исследовательский университет)»

Факультет информатики  
Кафедра технической кибернетики

**Лабораторная работа № 1  
по курсу «Интеллектуальный анализ данных»**

Студент:   
группа   
Преподаватель:

Самара 2018

Задание

1. Установить на компьютер программную систему статистического анализа и визуализации данных R (<http://www.r-project.org/>).

2. Установить интегрированную среду разработки RStudio (<http://www.rstudio.org/>).

3. Разобрать примеры использования различных функций R.

4. Подготовить текстовый файл данных для статистического анализа в системе R, выбрав таблицу и подходящие факторы из банка реальных данных в архиве. Выбрать пару факторов (столбцов таблицы Name\_K, Name\_M) для анализа коэффициента корреляции.

5. Для выбранных факторов рассчитать коэффициент корреляции, проверить гипотезу о некоррелированности выбранных факторов на уровне значимости 0.05.

6. Визуализировать точки исходных данных на плоскости (Name\_K, Name\_M).

7. Повторить расчёт коэффициента корреляции и визуализацию исходных данных для одного из наборов UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/).

Реферат

Отчёт по лабораторной работе 17 с, 1 рисунок, 0 таблиц, 2 приложения.

КОРРЕЛЯЦИЯ, СТАТИСТИКА, ЗАВИСИМОСТЬ, ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ, УРОВЕНЬ ЗНАЧИМОСТИ, КРИТЕРИЙ СТЬЮДЕНТА

В приведённой работе проводится исследование статистической зависимости между определёнными параметрами входного набора данных.

Цель работы – определить наличие корреляции между несколькими факторами, доверительный интервал и коэффициент корреляции для заданного уровня значимости.

Полученные результаты характеризуют наличие корреляции и отражают зависимость между выбранными параметрами. Присутствует графическое описание результатов эксперимента.

Работа была выполнена в среде моделирования R-Studio с применением скриптового языка программирования R.

Содержание

[Введение 5](#_Toc448163917)

[1 Корреляционный анализ данных 6](#_Toc448163918)

[2 Исходные данные 7](#_Toc448163919)

[3 Результаты 8](#_Toc448163920)

[Заключение 10](#_Toc448163921)

[Приложение А Набор данных 11](#_Toc448163922)

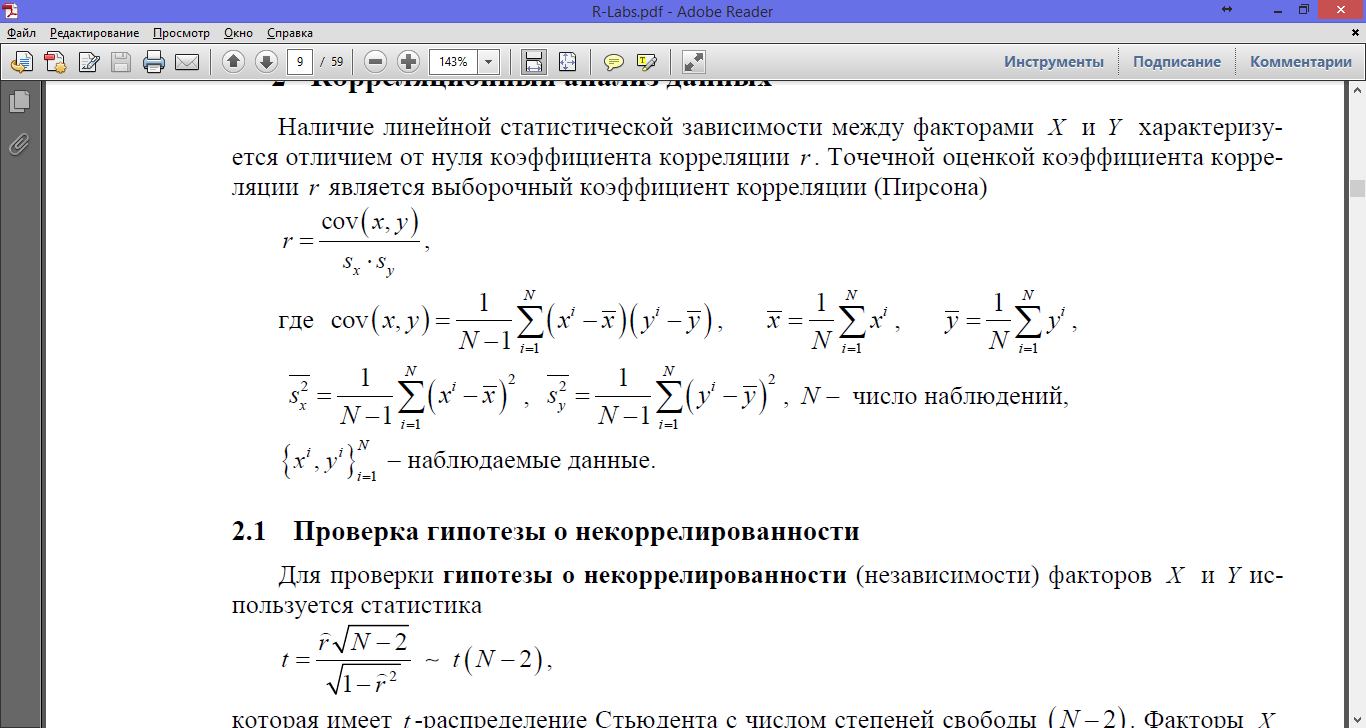
[Приложение Б Исходные коды программ 15](#_Toc448163923)

Введение

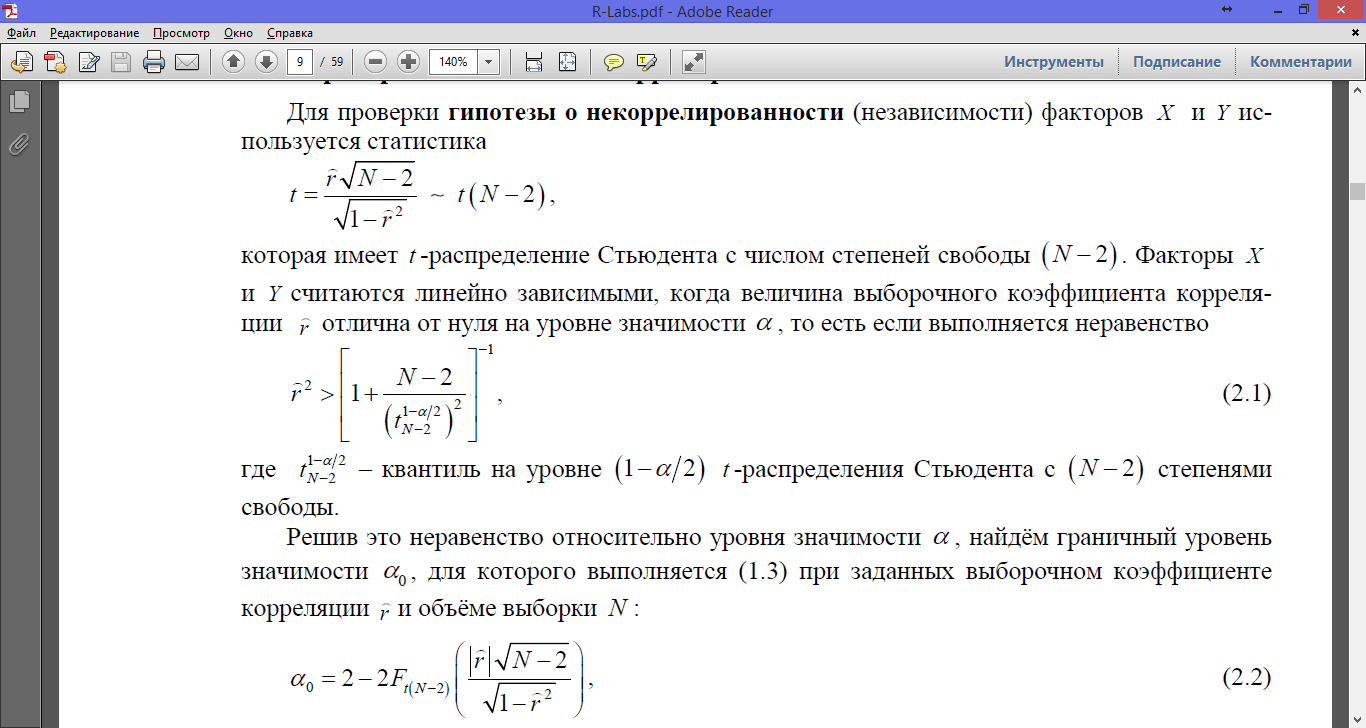
Существуют различные способы для проведения статистического анализа, начиная с ручных расчётов на бумаге и заканчивая применением современных вычислительных средств и технологий. В рамках решения поставленной задачи оптимальным способом является использование существующих пакетов для численных расчётов и анализа зависимостей, таких как MSExcelи RStudio. В нашем случае наиболее предпочтительным будет второй вариант, поскольку интегрированная среда разработки RStudioпредоставляет гораздо больше возможностей для статистической обработки данных.

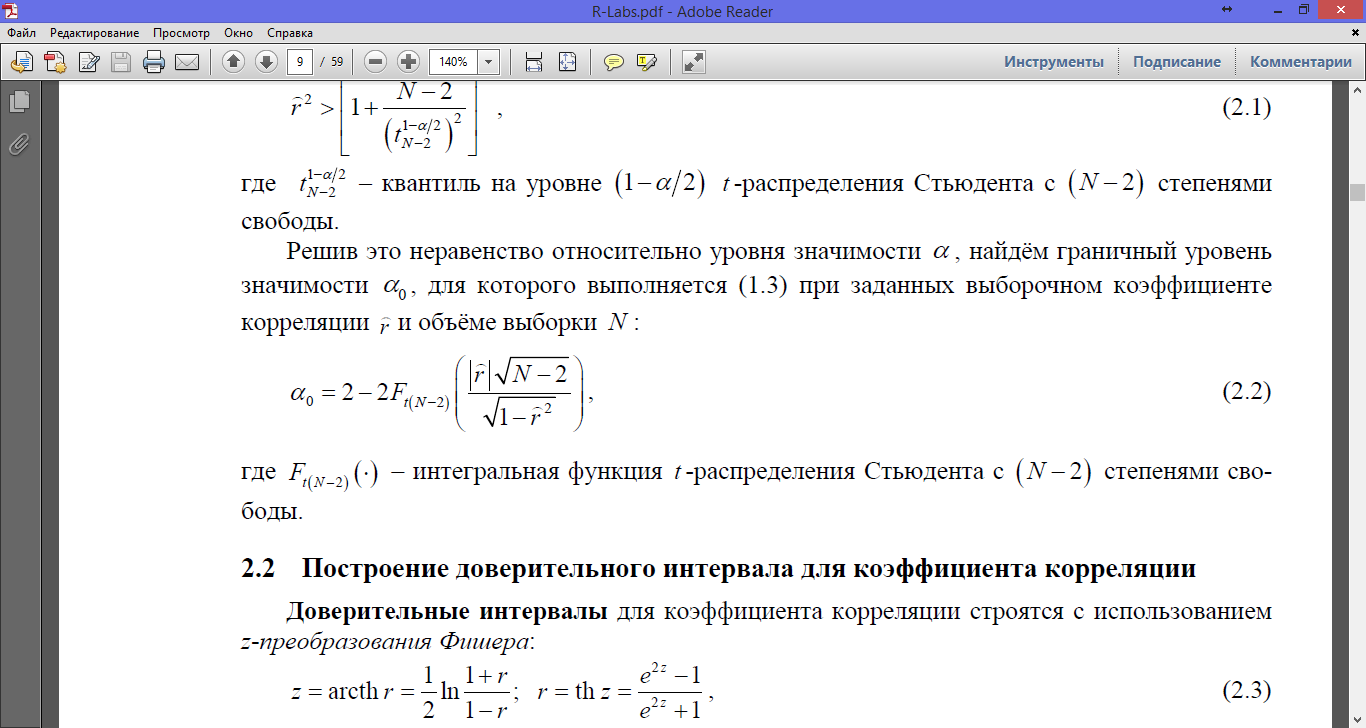
# 1

# Корреляционный анализ данных



Проверка гипотезы онекоррелированности





# 2 Исходные данные

В качестве исходных данных к заданию были взятыдва набора:

1. Олимпийское семиборье (Сеул, 1988)

Результаты всех 25 участниц

hurdles - бег с пепятствиями на 100 метров [c]

highjump - прыжки в высоту [м]

shot - бросание ядра [м]

run200m - бег на 200 метров [с]

longjump - прыжки в длину [м]

javelin - метанье копья [м]

run800m - бег на 800 метров [с]

score - сумма очков

1. IrisDataSet из репозитория UC Irvine Machine Learning–выборка из 150 экземпляров растений семейства ирисовые, содержащая набор факторов из пяти столбцов:

* sepal length in cm – длиначашелистика
* sepal width in cm – шириначашелистика
* petal length in cm – длиналепестка
* petal width in cm – шириналепестка
* class – наименование класса растения

На основании первых четырёх параметров определяется принадлежность конкретного экземпляра к одному из трёх классов, указанных в пятом столбце.

# 3 Результаты

Исходными данными для исследования коэффициента корреляции были выбраны:

Для первого набора – зависимость между результатами бега на 200 м и на 800 м;

Для второго набора – такие факторы как длина чашелистика (1 столбец) и длина лепестка (2 столбец) растений. Наборы данных были загружены из текстовых файлов в R-Studio и представлены в виде двумерных таблиц. Для анализа зависимости и графического представления результатов эксперимента использовались следующие функции:

cor.test(data$X1, data$X3);

plot(data$X1, data$X3);

Описание результата, полученного при выполнении указанного скрипта, приведено ниже:

Наборданных №1:

Pearson's product-moment correlation

data: test$200m and test$800m

t = 3.758191, df = 23, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.6084589 0.6250246

sample estimates:

cor

0.6168101

Здесь:

t – коэффициент t-статистики Стьюдента;

df – число степеней свободы;

p-value–граничный уровень значимости гипотезы о некоррелированности.

Как можно видеть, доверительный интервал для вероятности 0.95 лежит в пределах 0.608-0.625, следовательно, можно сделать вывод о наличии зависимости между выбранными факторами.

Наборданных №2:

Pearson's product-moment correlation

data: iris.data$X1 and iris.data$X3

t = 21.646, df = 148, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.8270368 0.9055083

sample estimates:

cor

0.8717542

Вывод аналогичный.

Графическая интерпретация полученных результатов представлена на рисунках 1 и 2.

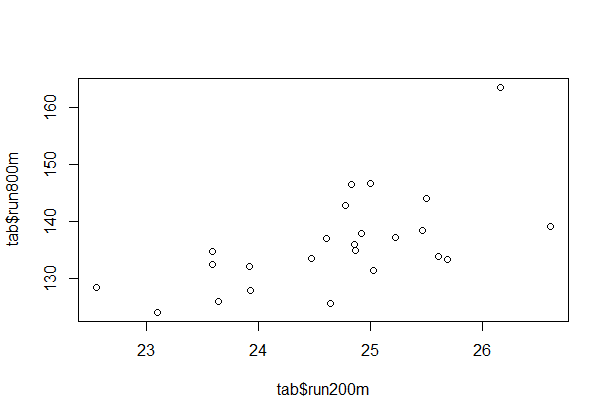


Рисунок 1 – статистическая зависимость времени бега на 200 м и на 800 м

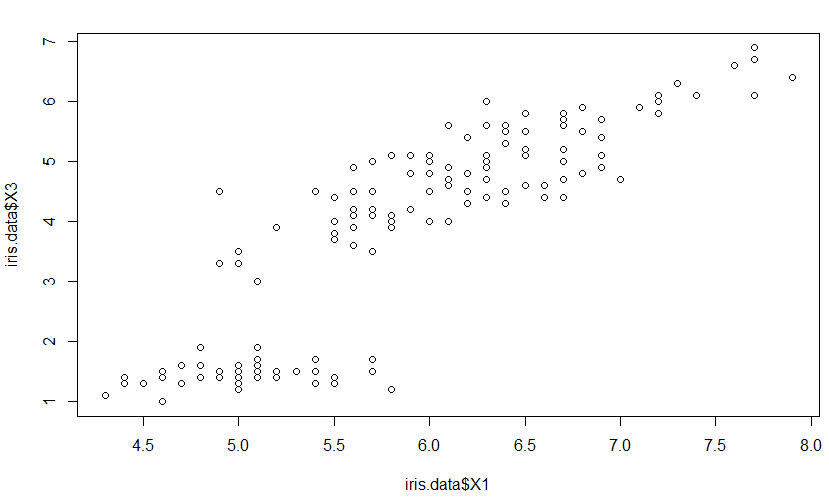


Рисунок 2 – статистическая зависимость длины чашелистика от длины лепестка ириса

Заключение

В процессе работы были проанализированы статистические данные из входного набора и выявлена статистическая зависимость между исследуемыми факторами. На основании полученных результатов можно сделать вывод о том, что между этими факторами присутствует корреляция на заданном уровне значимости.

Приложение А  
Набор данных

Набор данных №1:

---------------------------------------------------------------------------------

hurdles highjump shot run200m longjump javelin run800m score

---------------------------------------------------------------------------------

Joyner-Kersee (USA) 12.69 1.86 15.80 22.56 7.27 45.66 128.51 7291

John (GDR) 12.85 1.80 16.23 23.65 6.71 42.56 126.12 6897

Behmer (GDR) 13.20 1.83 14.20 23.10 6.68 44.54 124.20 6858

Sablovskaite (URS) 13.61 1.80 15.23 23.92 6.25 42.78 132.24 6540

Choubenkova (URS) 13.51 1.74 14.76 23.93 6.32 47.46 127.90 6540

Schulz (GDR) 13.75 1.83 13.50 24.65 6.33 42.82 125.79 6411

Fleming (AUS) 13.38 1.80 12.88 23.59 6.37 40.28 132.54 6351

Greiner (USA) 13.55 1.80 14.13 24.48 6.47 38.00 133.65 6297

Lajbnerova (CZE) 13.63 1.83 14.28 24.86 6.11 42.20 136.05 6252

Bouraga (URS) 13.25 1.77 12.62 23.59 6.28 39.06 134.74 6252

Wijnsma (HOL) 13.75 1.86 13.01 25.03 6.34 37.86 131.49 6205

Dimitrova (BUL) 13.24 1.80 12.88 23.59 6.37 40.28 132.54 6171

Scheider (SWI) 13.85 1.86 11.58 24.87 6.05 47.50 134.93 6137

Braun (FRG) 13.71 1.83 13.16 24.78 6.12 44.58 142.82 6109

Ruotsalainen (FIN) 13.79 1.80 12.32 24.61 6.08 45.44 137.06 6101

Yuping (CHN) 13.93 1.86 14.21 25.00 6.40 38.60 146.67 6087

Hagger (GB) 13.47 1.80 12.75 25.47 6.34 35.76 138.48 5975

Brown (USA) 14.07 1.83 12.69 24.83 6.13 44.34 146.43 5972

Mulliner (GB) 14.39 1.71 12.68 24.92 6.10 37.76 138.02 5746

Hautenauve (BEL) 14.04 1.77 11.81 25.61 5.99 35.68 133.90 5734

Kytola (FIN) 14.31 1.77 11.66 25.69 5.75 39.48 133.35 5686

Geremias (BRA) 14.23 1.71 12.95 25.50 5.50 39.64 144.02 5508

Hui-Ing (TAI) 14.85 1.68 10.00 25.23 5.47 39.14 137.30 5290

Jeong-Mi (KOR) 14.53 1.71 10.83 26.61 5.50 39.26 139.17 5289

Launa (PNG) 16.42 1.50 11.78 26.16 4.88 46.38 163.43 4566

---------------------------------------------------------------------------------

Набор данных №2:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **1** | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| **2** | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| **3** | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **4** | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| **5** | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| **6** | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| **7** | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| **8** | 5.0 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| **9** | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| **10** | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| **11** | 5.4 | 3.7 | 1.5 | 0.2 | Iris-setosa |
| **12** | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa |
| **13** | 4.8 | 3.0 | 1.4 | 0.1 | Iris-setosa |
| **14** | 4.3 | 3.0 | 1.1 | 0.1 | Iris-setosa |
| **15** | 5.8 | 4.0 | 1.2 | 0.2 | Iris-setosa |
| **16** | 5.7 | 4.4 | 1.5 | 0.4 | Iris-setosa |
| **17** | 5.4 | 3.9 | 1.3 | 0.4 | Iris-setosa |
| **18** | 5.1 | 3.5 | 1.4 | 0.3 | Iris-setosa |
| **19** | 5.7 | 3.8 | 1.7 | 0.3 | Iris-setosa |
| **20** | 5.1 | 3.8 | 1.5 | 0.3 | Iris-setosa |
| **21** | 5.4 | 3.4 | 1.7 | 0.2 | Iris-setosa |
| **22** | 5.1 | 3.7 | 1.5 | 0.4 | Iris-setosa |
| **23** | 4.6 | 3.6 | 1.0 | 0.2 | Iris-setosa |
| **24** | 5.1 | 3.3 | 1.7 | 0.5 | Iris-setosa |
| **25** | 4.8 | 3.4 | 1.9 | 0.2 | Iris-setosa |
| **26** | 5.0 | 3.0 | 1.6 | 0.2 | Iris-setosa |
| **27** | 5.0 | 3.4 | 1.6 | 0.4 | Iris-setosa |
| **28** | 5.2 | 3.5 | 1.5 | 0.2 | Iris-setosa |
| **29** | 5.2 | 3.4 | 1.4 | 0.2 | Iris-setosa |
| **30** | 4.7 | 3.2 | 1.6 | 0.2 | Iris-setosa |
| **31** | 4.8 | 3.1 | 1.6 | 0.2 | Iris-setosa |
| **32** | 5.4 | 3.4 | 1.5 | 0.4 | Iris-setosa |
| **33** | 5.2 | 4.1 | 1.5 | 0.1 | Iris-setosa |
| **34** | 5.5 | 4.2 | 1.4 | 0.2 | Iris-setosa |
| **35** | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| **36** | 5.0 | 3.2 | 1.2 | 0.2 | Iris-setosa |
| **37** | 5.5 | 3.5 | 1.3 | 0.2 | Iris-setosa |
| **38** | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| **39** | 4.4 | 3.0 | 1.3 | 0.2 | Iris-setosa |
| **40** | 5.1 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| **41** | 5.0 | 3.5 | 1.3 | 0.3 | Iris-setosa |
| **42** | 4.5 | 2.3 | 1.3 | 0.3 | Iris-setosa |
| **43** | 4.4 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **44** | 5.0 | 3.5 | 1.6 | 0.6 | Iris-setosa |
| **45** | 5.1 | 3.8 | 1.9 | 0.4 | Iris-setosa |
| **46** | 4.8 | 3.0 | 1.4 | 0.3 | Iris-setosa |
| **47** | 5.1 | 3.8 | 1.6 | 0.2 | Iris-setosa |
| **48** | 4.6 | 3.2 | 1.4 | 0.2 | Iris-setosa |
| **49** | 5.3 | 3.7 | 1.5 | 0.2 | Iris-setosa |
| **50** | 5.0 | 3.3 | 1.4 | 0.2 | Iris-setosa |
| **51** | 7.0 | 3.2 | 4.7 | 1.4 | Iris-versicolor |
| **52** | 6.4 | 3.2 | 4.5 | 1.5 | Iris-versicolor |
| **53** | 6.9 | 3.1 | 4.9 | 1.5 | Iris-versicolor |
| **54** | 5.5 | 2.3 | 4.0 | 1.3 | Iris-versicolor |
| **55** | 6.5 | 2.8 | 4.6 | 1.5 | Iris-versicolor |
| **56** | 5.7 | 2.8 | 4.5 | 1.3 | Iris-versicolor |
| **57** | 6.3 | 3.3 | 4.7 | 1.6 | Iris-versicolor |
| **58** | 4.9 | 2.4 | 3.3 | 1.0 | Iris-versicolor |
| **59** | 6.6 | 2.9 | 4.6 | 1.3 | Iris-versicolor |
| **60** | 5.2 | 2.7 | 3.9 | 1.4 | Iris-versicolor |
| **61** | 5.0 | 2.0 | 3.5 | 1.0 | Iris-versicolor |
| **62** | 5.9 | 3.0 | 4.2 | 1.5 | Iris-versicolor |
| **63** | 6.0 | 2.2 | 4.0 | 1.0 | Iris-versicolor |
| **64** | 6.1 | 2.9 | 4.7 | 1.4 | Iris-versicolor |
| **65** | 5.6 | 2.9 | 3.6 | 1.3 | Iris-versicolor |
| **66** | 6.7 | 3.1 | 4.4 | 1.4 | Iris-versicolor |
| **67** | 5.6 | 3.0 | 4.5 | 1.5 | Iris-versicolor |
| **68** | 5.8 | 2.7 | 4.1 | 1.0 | Iris-versicolor |
| **69** | 6.2 | 2.2 | 4.5 | 1.5 | Iris-versicolor |
| **70** | 5.6 | 2.5 | 3.9 | 1.1 | Iris-versicolor |
| **71** | 5.9 | 3.2 | 4.8 | 1.8 | Iris-versicolor |
| **72** | 6.1 | 2.8 | 4.0 | 1.3 | Iris-versicolor |
| **73** | 6.3 | 2.5 | 4.9 | 1.5 | Iris-versicolor |
| **74** | 6.1 | 2.8 | 4.7 | 1.2 | Iris-versicolor |
| **75** | 6.4 | 2.9 | 4.3 | 1.3 | Iris-versicolor |
| **76** | 6.6 | 3.0 | 4.4 | 1.4 | Iris-versicolor |
| **77** | 6.8 | 2.8 | 4.8 | 1.4 | Iris-versicolor |
| **78** | 6.7 | 3.0 | 5.0 | 1.7 | Iris-versicolor |
| **79** | 6.0 | 2.9 | 4.5 | 1.5 | Iris-versicolor |
| **80** | 5.7 | 2.6 | 3.5 | 1.0 | Iris-versicolor |
| **81** | 5.5 | 2.4 | 3.8 | 1.1 | Iris-versicolor |
| **82** | 5.5 | 2.4 | 3.7 | 1.0 | Iris-versicolor |
| **83** | 5.8 | 2.7 | 3.9 | 1.2 | Iris-versicolor |
| **84** | 6.0 | 2.7 | 5.1 | 1.6 | Iris-versicolor |
| **85** | 5.4 | 3.0 | 4.5 | 1.5 | Iris-versicolor |
| **86** | 6.0 | 3.4 | 4.5 | 1.6 | Iris-versicolor |
| **87** | 6.7 | 3.1 | 4.7 | 1.5 | Iris-versicolor |
| **88** | 6.3 | 2.3 | 4.4 | 1.3 | Iris-versicolor |
| **89** | 5.6 | 3.0 | 4.1 | 1.3 | Iris-versicolor |
| **90** | 5.5 | 2.5 | 4.0 | 1.3 | Iris-versicolor |
| **91** | 5.5 | 2.6 | 4.4 | 1.2 | Iris-versicolor |
| **92** | 6.1 | 3.0 | 4.6 | 1.4 | Iris-versicolor |
| **93** | 5.8 | 2.6 | 4.0 | 1.2 | Iris-versicolor |
| **94** | 5.0 | 2.3 | 3.3 | 1.0 | Iris-versicolor |
| **95** | 5.6 | 2.7 | 4.2 | 1.3 | Iris-versicolor |
| **96** | 5.7 | 3.0 | 4.2 | 1.2 | Iris-versicolor |
| **97** | 5.7 | 2.9 | 4.2 | 1.3 | Iris-versicolor |
| **98** | 6.2 | 2.9 | 4.3 | 1.3 | Iris-versicolor |
| **99** | 5.1 | 2.5 | 3.0 | 1.1 | Iris-versicolor |
| **100** | 5.7 | 2.8 | 4.1 | 1.3 | Iris-versicolor |
| **101** | 6.3 | 3.3 | 6.0 | 2.5 | Iris-virginica |
| **102** | 5.8 | 2.7 | 5.1 | 1.9 | Iris-virginica |
| **103** | 7.1 | 3.0 | 5.9 | 2.1 | Iris-virginica |
| **104** | 6.3 | 2.9 | 5.6 | 1.8 | Iris-virginica |
| **105** | 6.5 | 3.0 | 5.8 | 2.2 | Iris-virginica |
| **106** | 7.6 | 3.0 | 6.6 | 2.1 | Iris-virginica |
| **107** | 4.9 | 2.5 | 4.5 | 1.7 | Iris-virginica |
| **108** | 7.3 | 2.9 | 6.3 | 1.8 | Iris-virginica |
| **109** | 6.7 | 2.5 | 5.8 | 1.8 | Iris-virginica |
| **110** | 7.2 | 3.6 | 6.1 | 2.5 | Iris-virginica |
| **111** | 6.5 | 3.2 | 5.1 | 2.0 | Iris-virginica |
| **112** | 6.4 | 2.7 | 5.3 | 1.9 | Iris-virginica |
| **113** | 6.8 | 3.0 | 5.5 | 2.1 | Iris-virginica |
| **114** | 5.7 | 2.5 | 5.0 | 2.0 | Iris-virginica |
| **115** | 5.8 | 2.8 | 5.1 | 2.4 | Iris-virginica |
| **116** | 6.4 | 3.2 | 5.3 | 2.3 | Iris-virginica |
| **117** | 6.5 | 3.0 | 5.5 | 1.8 | Iris-virginica |
| **118** | 7.7 | 3.8 | 6.7 | 2.2 | Iris-virginica |
| **119** | 7.7 | 2.6 | 6.9 | 2.3 | Iris-virginica |
| **120** | 6.0 | 2.2 | 5.0 | 1.5 | Iris-virginica |
| **121** | 6.9 | 3.2 | 5.7 | 2.3 | Iris-virginica |
| **122** | 5.6 | 2.8 | 4.9 | 2.0 | Iris-virginica |
| **123** | 7.7 | 2.8 | 6.7 | 2.0 | Iris-virginica |
| **124** | 6.3 | 2.7 | 4.9 | 1.8 | Iris-virginica |
| **125** | 6.7 | 3.3 | 5.7 | 2.1 | Iris-virginica |
| **126** | 7.2 | 3.2 | 6.0 | 1.8 | Iris-virginica |
| **127** | 6.2 | 2.8 | 4.8 | 1.8 | Iris-virginica |
| **128** | 6.1 | 3.0 | 4.9 | 1.8 | Iris-virginica |
| **129** | 6.4 | 2.8 | 5.6 | 2.1 | Iris-virginica |
| **130** | 7.2 | 3.0 | 5.8 | 1.6 | Iris-virginica |
| **131** | 7.4 | 2.8 | 6.1 | 1.9 | Iris-virginica |
| **132** | 7.9 | 3.8 | 6.4 | 2.0 | Iris-virginica |
| **133** | 6.4 | 2.8 | 5.6 | 2.2 | Iris-virginica |
| **134** | 6.3 | 2.8 | 5.1 | 1.5 | Iris-virginica |
| **135** | 6.1 | 2.6 | 5.6 | 1.4 | Iris-virginica |
| **136** | 7.7 | 3.0 | 6.1 | 2.3 | Iris-virginica |
| **137** | 6.3 | 3.4 | 5.6 | 2.4 | Iris-virginica |
| **138** | 6.4 | 3.1 | 5.5 | 1.8 | Iris-virginica |
| **139** | 6.0 | 3.0 | 4.8 | 1.8 | Iris-virginica |
| **140** | 6.9 | 3.1 | 5.4 | 2.1 | Iris-virginica |
| **141** | 6.7 | 3.1 | 5.6 | 2.4 | Iris-virginica |
| **142** | 6.9 | 3.1 | 5.1 | 2.3 | Iris-virginica |
| **143** | 5.8 | 2.7 | 5.1 | 1.9 | Iris-virginica |
| **144** | 6.8 | 3.2 | 5.9 | 2.3 | Iris-virginica |
| **145** | 6.7 | 3.3 | 5.7 | 2.5 | Iris-virginica |
| **146** | 6.7 | 3.0 | 5.2 | 2.3 | Iris-virginica |
| **147** | 6.3 | 2.5 | 5.0 | 1.9 | Iris-virginica |
| **148** | 6.5 | 3.0 | 5.2 | 2.0 | Iris-virginica |
| **149** | 6.2 | 3.4 | 5.4 | 2.3 | Iris-virginica |
| **150** | 5.9 | 3.0 | 5.1 | 1.8 | Iris-virginica |

Приложение Б  
Исходные коды программ

№1

cor.test(test$wt, test$bmr);

plot(test$wt, test$bmr);

№2

cor.test(iris.data$X1, iris.data$X3);

plot(iris.data$X1, iris.data$X3);