Белорусский государственный университет
информатики и радиоэлектроники

Кафедра ЭВМ

Отчет по лабораторной работе №1
по курсу ХиУИ

Вариант #30

Выполнил:
студент группы 7М2432
Пантелеев В.В.

Проверил:
Марченко В.В.

МИНСК 2017

Цель работы:

1) Изучить методику проведения корреляционного анализа при помощи языка программирования R.

2) Провести статистическую оценку коэфицента Пирсона между двумя признаками.

ВВЕДЕНИЕ

Корреляционный анализ — метод обработки статистических данных, заключающийся в изучении коэффициентов (корреляции). Его применение возможно в случае наличия достаточного количества (для конкретного вида коэффициента корреляции) наблюдений из более чем одной переменной. При этом сравниваются коэффициенты корреляции между одной парой или множеством пар признаков, для установления между ними статистических взаимосвязей. Данный метод обработки статистических данных весьма популярен в социальных науках (в частности в психологии), хотя сфера применения коэффициентов корреляции обширна: контроль качества промышленной продукции, металловедение, агрохимия и проч. Популярность метода обусловлена двумя моментами: коэффициенты корреляции относительно просты в подсчете, их применение не требует специальной математической подготовки. В сочетании с простотой интерпретации (принятие гипотезы о наличии корреляции означает что изменение переменной А, произойдет одновременно с изменением значения Б), простота применения коэффициента привела к его широкому распространению в сфере анализа статистических данных.

Выполнение работы

Входные данные: n объектов, каждый из которых характеризуется двумя числовыми признаками: X, Y.

Требуется исследовать степень взаимосвязи между двумя признаками некоторых объектов. Для каждого набора данных необходимо выполнить следующие задания.

1. Визуализировать данные на плоскости в виде точек с координатами {x, y}

2. Статистически оценить коэффициент корреляции Пирсона между признаками x и y.

3. Проверить статистическую гипотезу о некоррелированности признаков x и y на уровне значимости 0,05.

Все описанные выше задания требуется выполнить для двух наборов данных.

1. Смоделированные независимые случайные векторы (X , Y), имеющие гауссовское распределение с заданным математическим ожиданием a и корреляционной матрицей R.

2. Реальные статистические данные из заданного набора (выдаются преподавателем).

Исходные данные — n = 10000, a = (1, -1), R = $\begin{matrix} 16, -15 \\ -15, 16 \end{matrix}$

Листинг программы:

```
require(MASS)

n <- 10000
a <- c(1, -1)
r <- cbind(c(16, -15), c(-15, 16))

analyse_cor <- function(x, y) {
  print(cor.test(x, y))
  dev.new()
  plot(x, y)
}
```

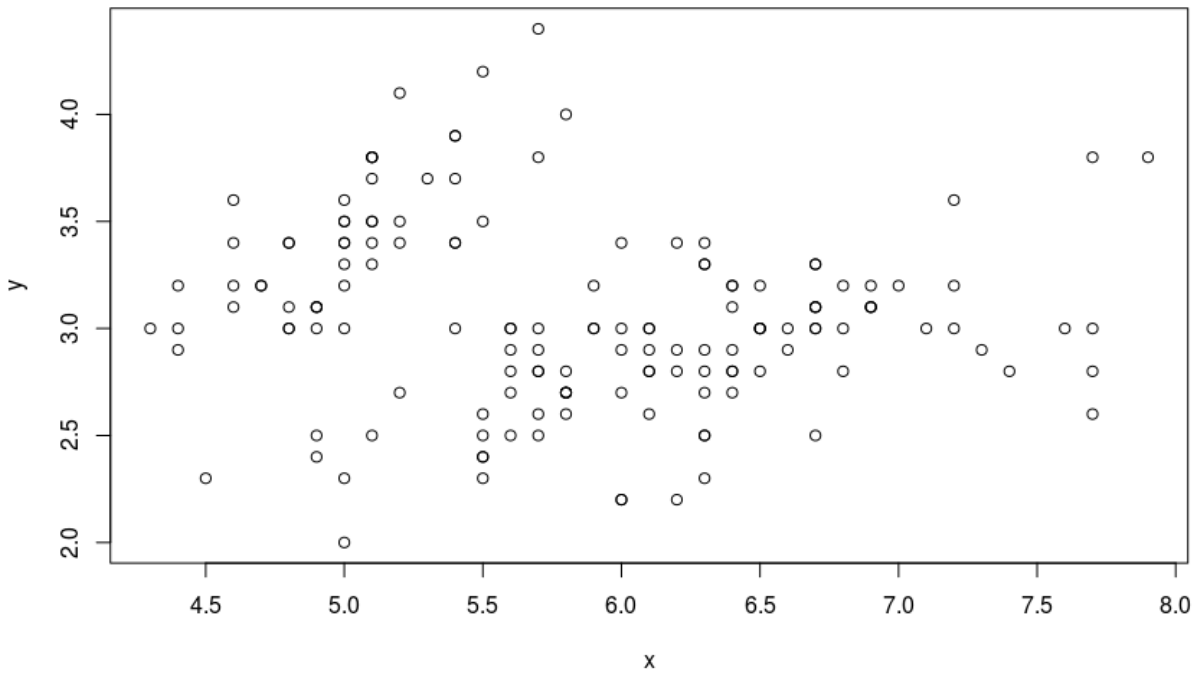
```
fileName = "/home/panteleev/Documents/Data/iris.data.txt"
irisData <- read.table(fileName, header = FALSE, sep = ",")

head(irisData)

analyse_cor(irisData$V1, irisData$V2)

samopalData <- mvrnorm(n, a, r)

head(samopalData)

analyse_cor(samopalData[,1], samopalData[,2])
```

Результат выполнения программы:

Часть загруженных данных из файла *iris.data.txt:*

```
  V1  V2  V3  V4          V5
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
```

Визуализация данных из файла *iris.data.txt:*



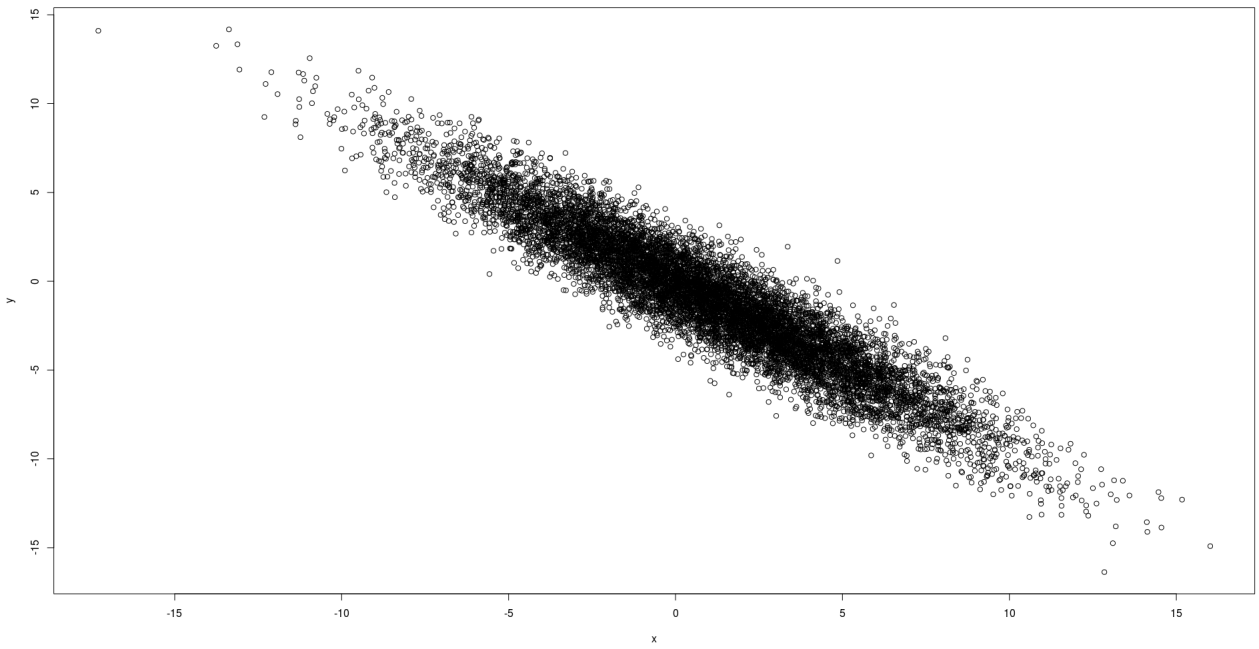
Результат анализа при помощи коэфицента Пирсона данных из файла *iris.data.txt:*

```
        Pearson's product-moment correlation

data:  x and y
t = -1.3386, df = 148, p-value = 0.1828
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.26498618  0.05180021
sample estimates:
        cor
-0.1093692
```

Часть сгенерированных данных*:*

```
           [,1]      [,2]
[1,]  0.9437493 -2.043852
[2,]  0.2879423  0.820856
[3,]  6.4675847 -7.542139
[4,]  4.1195804 -2.944259
[5,] -4.0613449  4.527895
[6,] -2.4062410  2.930632
```

Визуализация сгенерированных данных:



Результат анализа при помощи коэфицента Пирсона сгенерированных данных *:*

```
        Pearson's product-moment correlation

data:  x and y
t = -273.63, df = 9998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9415195 -0.9369000
sample estimates:
        cor
-0.9392523
```

Выводы

Для данных из файла *iris.data.txt,* значение коэфицента корелляции ровняется ~ -0.1, для сгенерированных данных коэфицент корелляции ровняется ~ -0.93. Можно сказать что данные загруженные из файла практически не коррелируют, что отчетливо видно на графики визуализации данных, с другой стороны сгенерированные данные коррелируют очень сильно, что можно так же отчетливо увидеть на графике.