Белорусский Государственный Университет
Информатики и Радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра ЭВМ

Лабораторная работа №1

Тема «Корреляционный анализ»

Выполнил:

Студент группы 7М2432

Пашковский А.А.

Проверил:

Марченко В.В.

Минск, 2017

Задание:

Входные данные: n объектов, каждый из которых характеризуется двумя числовыми признаками: $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^n$.

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 $\{(x_i, y_i)\}_{i=1}^n$.

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.

Исходные данные:

1) значения объёма исследуемой выборки (n) – 1000;

2) значения вектора математических ожиданий (a) – (-1, 0);

3) корреляционных матриц (R) для моделируемой выборки из гауссовских случайных векторов –$\begin{pmatrix} 4 & -3 \\ -3 & 9 \end{pmatrix}$

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).

26. Parkinsons Disease Data Set

Название файла: `26-parkinsons.txt`

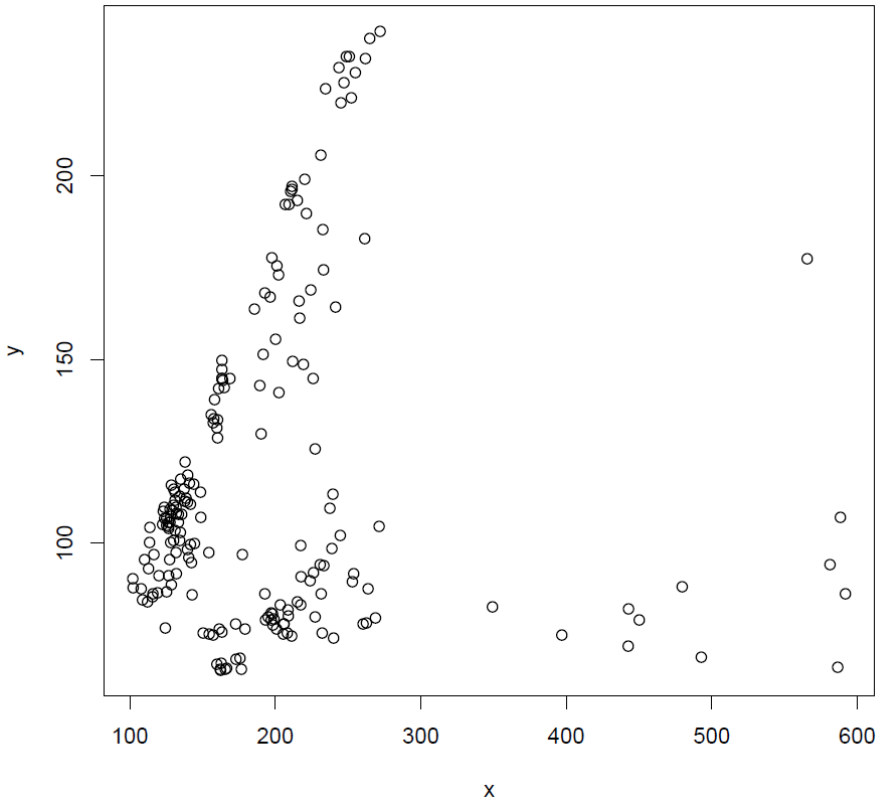
Ссылка: http://archive.ics.uci.edu/ml/datasets/Parkinsons

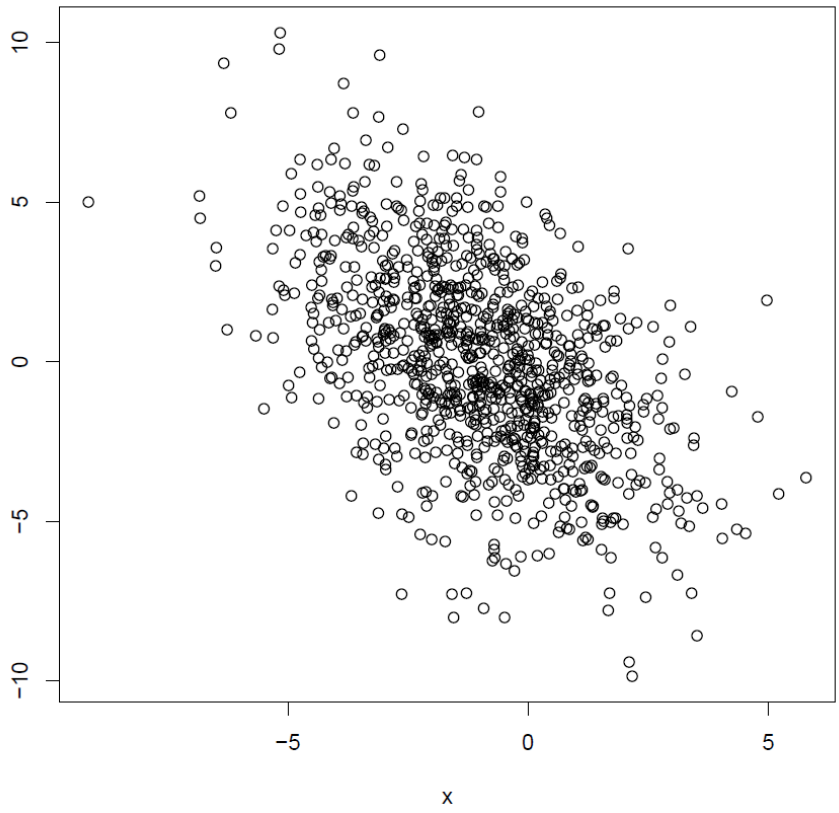
Первый признак: MDVP:Fhi(Hz) (столбец № 3)

Второй признак: MDVP:Flo(Hz) (столбец № 4)

Результаты:

1. Изображения данных в виде точек на плоскости:





2. Статистические оценки коэффициентов корреляции Пирсона для каждого набора данных, сравнение статистической оценки коэффициента корреляции Пирсона с реальным коэффициентом корреляции Пирсона для смоделированных данных:

a) Данные из parkinsons.data.txt:

```
data:  x and y
t = 1.1845, df = 193, p-value = 0.2377
alternative hypothesis: true correlation is not equal to 0
5 percent confidence interval:
 0.08045674 0.08944231
sample estimates:
       cor
0.08495125
```

Число стьюдента для уровня значимости 0.05 и степеней свободы 200 равно 1.971

Т.к. |t| << 1,971, то гипотеза о некоррелированности принимается.

b) Данные из модуляции по выборке:

```
data:  x and y
t = -18.621, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
5 percent confidence interval:
 -0.5092568 -0.5063090
sample estimates:
       cor
-0.5077844
```

Число Стьюдента для уровня значимости 0,05 и степеней свободы >100 равно 1,96.

Т.к. |t| > 1,96, то гипотеза о некоррелированности отвергается.

Листинг программы:

```r
require(MASS)

analyse_cor <- function(x, y) {
  print(cor.test(x, y, conf.level = 0.05))
  dev.new()
  plot(x, y)
}

dat <- read.table("parkinsons.data.txt", sep=",")
analyse_cor(dat$V3, dat$V4)

n <- 1000
a <- c(-1, 0)
r <- cbind(c(4, -3), c(-3, 9))
dat <- mvrnorm(n, a, r)
analyse_cor(dat[,1], dat[,2])
```