МЕЖГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Автоматизированные системы управления»

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Методические рекомендации к лабораторным работам для студентов специальности 1-53 01 02 «Автоматизированные системы обработки информации» очной и заочной формы обучения

Могилев 2022

УДК 519.6
ББК 22.176
С48

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Автоматизированные системы управления» «11» января 2022 г., протокол № 6

Составители: д-р техн. наук, доц. А. И. Якимов;
канд. техн. наук Е. А. Якимов

Рецензент канд. техн. наук, доц. С. К. Крутолевич

Изложены рекомендации по выполнению лабораторных работ, приведены примеры решения заданий, а также учебно-методическая литература. Предназначены для самостоятельного обучения при подготовке к выполнению лабораторных работ студентами специальности 1-53 01 02 «Автоматизированные системы обработки информации» очной и заочной формы обучения.

Учебно-методическое издание

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Ответственный за выпуск	А. И. Якимов
Корректор	Т. А. Рыжикова
Компьютерная верстка	Н. П. Полевничая

Подписано в печать . Формат 60×84/16. Бумага офсетная. Гарнитура Таймс. Печать трафаретная. Усл. печ. л. . Уч.-изд. л. . Тираж 26 экз. Заказ № .

Издатель и полиграфическое исполнение:
Межгосударственное образовательное учреждение высшего образования
«Белорусско-Российский университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/156 от 07.03.2019.
Пр-т Мира, 43, 212022, г. Могилев.

Содержание

Введение

Целью преподавания дисциплины «Случайные процессы и статистические методы обработки данных» является обучение студентов основным математическим моделям случайных процессов и статистическим методам обработки данных для решения задач из области программирования, администрирования сетей, информационных потоков, планирования, проектирования и управления автоматизированными системами.

Цель методических рекомендаций – помочь студентам в самостоятельной подготовке к выполнению заданий для лабораторных занятий по дисциплине.

Порядок выполнения каждой лабораторной работы.

1 Изучить теоретические сведения.

2 Получить задание у преподавателя, выполнить в соответствии с заданным вариантом.

3 Сделать выводы по результатам выполнения задания.

4 Оформить отчет.

Требования к отчету:

1 Цель работы.

2 Постановка задачи.

3 Результаты выполнения задания.

4 Выводы.

1 ЛАБОРАТОРНАЯ РАБОТА № 1. Нахождение описательной статистики экспериментальных данных средствами среды программирования R

Цель работы: Изучение описательной статистики экспериментальных данных средствами среды программирования R.

Основные теоретические положения

R - бесплатное свободно распространяемое программное обеспечение с открытым исходным кодом. Находит широкое применение в различных областях знаний для моделирования, статистического анализа и обработки данных.

Основные достоинства:

- Быстродействие.
- Гибкость.
- Разнообразие существующих пакетов, расширяющих базовый функционал.
- Кроссплатформенность.
- Активное сообщество.

Для работы с R нам потребуются:

- Язык R (https://cran.r-project.org/bin/windows/base/)
- Среда разработки RStudio (https://www.rstudio.com/products/rstudio/download/)
- Учебные материалы с сайта http://www.qsar4u.com/files/rintro/01.html

R - язык функционального программирования

Функции производят операции над объектом и возвращают результат, при этом передаваемый объект не изменяется.

\[result = f(data) \]

Если необходимо изменить состояние объекта, то результат функции присваивается переменной обозначающей этот объект.

\[data = f(data) \]

Любая операция в R это функция

```
2 + 3
```

```
[1] 5
```

```
`+` (2, 3)
```

```
[1] 5
```

Типы данных

Типы данных в порядке увеличения приоритета:

- Логические (logical)
- Целочисленные (integer)
- Вещественные числа (numeric)

- Комлексные числа (complex)
- Текстовые (character)
- Списки (list)

Векторы и типы данных

Вектор может содержать данные только одного типа.
Логические:

```
c(TRUE, TRUE, FALSE)
[1]  TRUE  TRUE FALSE
class(c(TRUE, TRUE, FALSE))
[1] "logical"
```

Целочисленные:

```
c(1, 2, TRUE)
[1] 1 2 1
class(c(1, 2, TRUE))
[1] "numeric"
```

Текстовые:

```
c(2, "boo")
[1] "2"   "boo"
class(c(2, "boo"))
[1] "character"
```

Способы создания векторов

```
c(1,3,7)
[1] 1 3 7
1:15
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
15:1
 [1] 15 14 13 12 11 10  9  8  7  6  5  4  3  2  1
-10:-1
 [1] -10  -9  -8  -7  -6  -5  -4  -3  -2  -1
-3.8:10
 [1] -3.8 -2.8 -1.8 -0.8  0.2  1.2  2.2  3.2  4.2  5.2  6.2  7.2
8.2  9.2
seq(from=2, to=20, by=2)
 [1]  2  4  6  8 10 12 14 16 18 20
seq(2, 20, 2)
 [1]  2  4  6  8 10 12 14 16 18 20
c(1:5, 8, 11:15)
 [1]  1  2  3  4  5  8 11 12 13 14 15
c(1:5, seq(10, 20, 2))
 [1]  1  2  3  4  5 10 12 14 16 18 20
```

Матрицы

Матрица - двумерный набор элементов *одного* типа (таблица).

```
matrix(data=1:12, nrow=3, ncol=4)
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

```
matrix(data=1:12, nrow=3, ncol=4, byrow=TRUE)
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
```

Массивы

Массив - многомерный набор элементов *одного* типа.

```
array(data=1:12, dim=c(2,2,3))
, , 1

     [,1] [,2]
[1,]    1    3
[2,]    2    4

, , 2

     [,1] [,2]
[1,]    5    7
[2,]    6    8

, , 3

     [,1] [,2]
[1,]    9   11
[2,]   10   12
```

Factors

Factor - представляет номинальную или ранговую шкалу. Используется для представления Y в классификационных моделях.

```
factor(c(1,1,0,0,1,1,0))
[1] 1 1 0 0 1 1 0
Levels: 0 1
a <- factor(c("active", "inactive", "active", "active", "inac-
tive", "active"))
a
[1] active   inactive active   active   inactive active
Levels: active inactive
levels(a)
[1] "active"   "inactive"
```

Data.frames

Data.frame - двумерный набор данных (таблица). В отличие от матриц, колонки в data.frame могут содержать данные различного типа. Однако тип данных внутри каждой колонки может быть только один. Это объясняется тем, что data.frame это список векторов (колонок). Поэтому к data.frame могут быть применены различные функции применимые к спискам.

```
data.frame(name=c("Tom", "Cruz", "Angela"), eye=c("brown",
"black", "green"),
           height=c(180, 192, 178))
    name   eye height
```

```
1     Tom brown    180
2    Cruz black    192
3 Angela green    178
```

Формулы

Формулы - специальная форма выражения отношений между переменными в уравнении. Формулы используются при построении моделей для определения функциональной зависимости между параметрами.

Линейная комбинация (+):

```
# y = a*x1 + b*x2 + c
formula("y ~ x1 + x2")
y ~ x1 + x2
```

Линейная комбинация с отсутствующим свободным членом (**+0**)

```
# y = a*x1 + b*x2
formula("y ~ x1 + x2 + 0")
y ~ x1 + x2 + 0
```

Функция идентичности **I()**, при этом выражение в скобках рассматривается как обычное математическое.

```
# y = a*(x1 + x2) + c
formula("y ~ I(x1 + x2)")
y ~ I(x1 + x2)
# y = a*x + b*x^2 + c
formula("y ~ x + I(x^2)")
y ~ x + I(x^2)
```

Формулы могут содержать математические функции

```
# log(y) = a*log(x1) + b*x2 + c
```

Примеры формул		
Синтаксис	Модель	Пояснение
Y ~ A	\(Y = \beta_{0} + \beta_{1}A \)	Уравнение регрессии с неявно заданным свободным членом
Y ~ A + 0	\(Y = \beta_{1}A \)	Уравнение регрессии без свободного члена
Y ~ A + B	\(Y = \beta_{0} + \beta_{1}A + \beta_{2}B \)	Уравнение модели первого порядка
Y ~ A + I(A^2)	\(Y = \beta_{0} + \beta_{1}A + \beta_{2}A^2 \)	Уравнение модели второго порядка с одной переменной
Y ~ A:B	\(Y = \beta_{0} + \beta_{1}AB \)	Уравнение модели первого порядка, в которое входят только произведения переменных
Y ~ A*B	\(Y = \beta_{0} + \beta_{1}A + \beta_{2}B + \beta_{3}AB \)	Полное уравнение модели первого порядка, аналогично Y ~ A + B + A:B
Y ~ (A + B + C)^2	\(Y = \beta_{0} + \beta_{1}A + \beta_{2}B + \beta_{3}C + \beta_{4}AB + \beta_{5}AC + \beta_{6}BC \)	Модель первого порядка включающая все произведения до порядка n, аналогично Y ~ A*B*C - A:B:C

```
formula("log(y) ~ log(x1) + x2")
log(y) ~ log(x1) + x2
```

Символ точки (.) подставляет все имеющиеся переменные. Функция зависимости y от всех остальных переменных, которые будут передаваться в функцию выглядит так

```
formula(y ~ .)
y ~ .
```

Списки

Списки содержат упорядоченный набор элементов, каждый из которых может являться вектором, матрицей, массивом, списком и т.д.

```
list(element_one=1:10, element_two=c(1, 3:7))
$element one
 [1]  1  2  3  4  5  6  7  8  9 10

$element two
[1] 1 3 4 5 6 7
list(1:10, c(1, 3:7))
[[1]]
 [1]  1  2  3  4  5  6  7  8  9 10

[[2]]
[1] 1 3 4 5 6 7
list(element_one=1:10, element_two=matrix(1:8, 2, 4))
$element_one
 [1]  1  2  3  4  5  6  7  8  9 10

$element_two
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
```

Как правило, в виде списков удобно хранить либо какие-то однотипные данные соответствующие разным итерациям, например, множество моделей. Или хранить разнородные данные, которые имеют смысловую связь, например, различные статистические характеристики отдельной модели.

Векторизация

В R действия выполняются **поэлементно** сразу над всем набором данных (будь то вектор, матрица, data.frame и т.д.). Векторные вычисления очень эффективны, поэтому всегда когда нужно совершить действия над элементами вектора (матрицы и т.д.) предпочтительнее использовать векторизацию, а не циклы.

```
a <- 1:10
a
 [1]  1  2  3  4  5  6  7  8  9 10
a + 2
 [1]  3  4  5  6  7  8  9 10 11 12
```

Кстати оператор присваивания тоже является функцией, и присвоение можно выполнить в таком виде

```
'<-'(aa, 1:10)
aa
 [1]  1  2  3  4  5  6  7  8  9 10
```

Сложение двух векторов одинаковой длины происходит поэлементно

```
a <- 1:10
a
 [1]  1  2  3  4  5  6  7  8  9 10
b <- 11:20
b
 [1] 11 12 13 14 15 16 17 18 19 20
a + b
 [1] 12 14 16 18 20 22 24 26 28 30
```

Как упоминалось, векторные вычисления могут производиться над любой структурой данных (вектор, матрица, data.frame и т.д.). Пусть у нас есть матрица

```
m <- matrix(1:12, nrow=4)
m
     [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
```

Умножим все ее элементы на 2, или возведем во вторую степень

```
m * 2
     [,1] [,2] [,3]
[1,]    2   10   18
[2,]    4   12   20
[3,]    6   14   22
[4,]    8   16   24
m ^ 2
     [,1] [,2] [,3]
[1,]    1   25   81
[2,]    4   36  100
[3,]    9   49  121
[4,]   16   64  144
```

Особенности векторизации

Если два вектора имеют различное количество элементов, то вектор меньшей длины будет повторяться столько раз чтобы соответствовать длине большего вектора.

Если длина большего вектора кратна длине меньшего вектора, такая операция будет произведена неявно, без уведомления пользователя.

```
a <- 1:10
a
 [1]  1  2  3  4  5  6  7  8  9 10
b <- 1:5
b
[1] 1 2 3 4 5
a + b
 [1]  2  4  6  8 10  7  9 11 13 15
```

В случае если длины комбинируемых векторов различны, то будет произведено циклическое совмещение элементов меньшего вектора относительно элементов большего вектора и будет сгенерировано предупреждение.

```
a <- 1:10
b <- 1:3
a + b
Warning: длина большего объекта не является произведением длины
меньшего
объекта
 [1]  2  4  6  5  7  9  8 10 12 11
```

Имена элементов векторов, матриц, data.frames, списков и т.д.

Все объекты поддерживают присвоение имен содержащимся в них элементам.

Обычный и именованный вектор

```
a <- 1:5
a
[1] 1 2 3 4 5
names(a) <- c("one", "two", "three", "four", "five")
a
  one   two three  four  five
    1     2     3     4     5
```

Другой способ создания именованного вектора

```
a <- c("one"=1, "two"=2, "three"=3, "four"=4, "five"=5)
a
  one   two three  four  five
    1     2     3     4     5
```

Аналогично векторам матрицы и data.frames имеют такие свойства как rownames и colnames, которые позволяют изменять имена колонок и строк.

```
df <- data.frame(1:2, 3:4, 5:6, 7:8)
df
  X1.2 X3.4 X5.6 X7.8
1    1    3    5    7
2    2    4    6    8
rownames(df) <- c("case_1", "case_2")
colnames(df) <- c("var_1", "var_2", "var_3", "var_4")
df
       var_1 var_2 var_3 var_4
case_1     1     3     5     7
case_2     2     4     6     8
```

Удаление имен осуществляется присвоением специального типа NULL

```
colnames(df) <- NULL
df
       NA NA NA NA
case_1  1  3  5  7
case_2  2  4  6  8
```

Или для векторов

```
a <- 1:5
```

```
names(a) <- c("one", "two", "three", "four", "five")
a
  one   two three  four  five
    1     2     3     4     5
names(a) <- NULL
a
[1] 1 2 3 4 5
```

При этом к элементам уже нельзя будет обращаться по имени, а только по индексу.

Преобразование типов и структур данных друг в друга

Преобразование типов данных осуществляется через группу функций, начинающихся на as.

Пример конвертации целочисленного вектора в текстовый

```
a <- 1:10
a
 [1]  1  2  3  4  5  6  7  8  9 10
as.character(a)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
```

Особенности приведения чисел, выраженных как factors к числовому виду.

Преобразуем вектор целых чисел в номинальную шкалу (factor).

```
a <- c(1,1,0,0,1,1,0)
a
[1] 1 1 0 0 1 1 0
a <- as.factor(a)
a
[1] 1 1 0 0 1 1 0
Levels: 0 1
```

Для обратной конвертации использование функции as.integer недостаточно.

```
as.integer(a)
[1] 2 2 1 1 2 2 1
```

Это связано с внутренним представлением типа данных factor. Эта структура представляет собой набор целочисленных значений, каждому из которых присвоено имя. В данном примере именами являются значения 1 и 0. Поэтому для корректного преобразования factor в целочисленный тип данных необходимо предварительно провести конвертацию в текстовый вид.

```
as.integer(as.character(a))
[1] 1 1 0 0 1 1 0
```

Данная операция часто вызывает затруднение и служит причиной ошибок.

Подобно преобразованию типов данных возможно **приведение различных структур данных к другому типу** с помощью того же семейства функций, начинающихся на as.

```
a <- 1:3
as.data.frame(a)
  a
1 1
2 2
3 3
```

Преобразоваание матрицы в data.frame

```
m <- matrix(1:8, 2, 4)
m
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
as.data.frame(m)
  V1 V2 V3 V4
1  1  3  5  7
2  2  4  6  8
```

Выборочной средней $x_{ср}$ называют среднее арифметическое значение признака выборочной совокупности. Если все значения x_1, x_2, …,x_n признака выборки объема n различны, то

$$\bar{x} = (x_1 + x_2 + ... + x_n)/n.$$

Описание функции

`mean(x, ...)`

Параметры

`x`	Вектор, матрица или data.frame.

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> mean(x)
7.12 (значение среднего)
```

Выборочной дисперсией называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $x_{ср}$. Если в качестве оценки генеральной дисперсии принять выборочную дисперсию, то эта оценка будет приводить к систематическим ошибкам, давая заниженное значение генеральной дисперсии, поэтому используют исправленную дисперсию. Исправленная (несмещенная) выборочная дисперсия вычисляется по формуле

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Выборочное среднеквадратическое отклонение - СКО $S_x = \sqrt{S_x^2}$

Описание функции

```
var(x, y = NULL, na.rm = FALSE)
sd(x, na.rm = FALSE)
```

Параметры

x	Вектор, матрица или data.frame
y	NULL(по умолчанию) или вектор, матрица или data.frame такой же размерности, что и x
na.rm	Удалить данные, значения котрорых отсуствуют

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.9,6.5)
> var(x, y, na.rm = FALSE)
> sd(x, na.rm = FALSE)

[1] 2.9       (значение дисперсии)
[1] 2.459065  (значение СКО)
```

Медианой m_e называют варианту, которая делит вариационный ряд (упорядоченный по возрастанию) на две части, равные по числу вариант.

Если число вариант нечетно, т.е. $n = 2r+1$, то $m_e=x_{r+1}$;

при четном $n = 2r$, то me=(xr+xr+1)/2

Модой M0 называют варианту, которая имеет наибольшую частоту. В R для этого можно использовать построение таблицы вариант

Описание функции

```
median(x, na.rm=FALSE)
```

Параметры

x Вектор, матрица или data.frame

na.rm Удалить отсутствующие данные.

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9,9.6,5.7,8.9,9.6)
> median(x, na.rm=FALSE)
8.9

> sort(unique(x))
3.6 7.8 9.6 5.7 8.9  уникальные значения вариант по возрастанию

> x.t<-table(x) > x.t
3.6 5.7 7.8 8.9 9.6          варианты (по возрастанию)
1   2   1   2   3            сколько раз каждая встречалась

> order(x.t)
1 3 2 4 5              номера вариант
(по частоте встречаемости)

> sort(unique(x))[which.max(x.t)]
9.6                          мода
```

Пакет MS Excel

Пакет MS Excel оснащен средствами статистической обработки данных. И хотя Excel существенно уступает специализированным статистическим пакетам обработки данных, тем не менее этот раздел математики представлен в Excel наиболее полно. В него включены основные, часто используемые статистические процедуры: средства описательной статистики, критерии различия, корреляционные и другие методы, позволяющие проводить необходимый статистический анализ экономических, медико-биологических и иных типов данных.

При рассмотрении применения методов обработки статистических данных ограничимся только простейшими и наиболее часто используемыми методами, реализованными в **Мастере функций** и **Пакете анализа**.

Выборочный метод.

По охвату статистической совокупности исследование может быть сплошное или не сплошное. При сплошном статистическом исследовании группа наблюдения формируется путем полного охвата всех единиц изучаемого явления. Множество всех единиц наблюдения, охватываемых таким сплошным наблюдением, называется генеральной совокупностью.

Основным методом не сплошного наблюдения является выборочный метод. Выборка – это группа элементов, выбранная для исследования из всей совокупности элементов. Конечной целью изучения выборочной совокупности всегда является получение информации о генеральной совокупности. Поэтому естественно стремиться сделать выборку так, чтобы она наилучшим образом представляла всю генеральную совокупность, то есть была бы репрезентативной или представительной.

Выборочная функция распределения.

В практических задачах закон распределения случайных величин обычно неизвестен или известен с точностью до некоторых неизвестных параметров. В частности, невозможно рассчитать точное значение соответствующих вероятностей, так как нельзя определить количество общих и благоприятных исходов. Поэтому вводится статистическое определение вероятности. По этому определению вероятность равна отношению числа испытаний, в которых событие появилось, к общему количеству произведенных испытаний. Такая вероятность называется статистической частотой.

В Excel для построения выборочных функций распределения используются специальная функция **Частота** и процедура Пакета анализа **Гистограмма**. Функция **Частота** вычисляет частоты появления случайной величины в интервалах значений и выводит их как массив чисел. Функция задается в качестве формулы массива. Синтаксис:

Частота (массив_данных; массив карманов),

где массив_данных – это массив или ссылка на множество данных, для которых вычисляются частоты; массив карманов – это массив или ссылка на множество интервалов, в которые группируются значения аргумента массив_данных.

Процедура **Гистограмма** используется для вычисления выборочных и интегральных частот попадания данных в указанные интервалы значений. Процедура выводит результаты в виде таблицы и гистограммы.

Задания для лабораторной работы

1. Имеются данные о цене на нефть x (ден. ед.) и индексе акций нефтяных компаний y (усл. ед.).

Вариант №	Данные						
1	*x*	18,26	18,06	19,40	19,70	20,10	19,60
	y	567	564	570	575	580	572
2	*x*	18,71	19,44	18,65	17,18	17,77	17,39
	y	556,3	555,94	544,9	540,72	541,07	559,47
3	*x*	18,99	17,71	18,82	17,42	18,4	18,15
	y	549,66	555,85	541,18	544,68	537,08	538,97
4	*x*	19,11	18,72	18,56	18,99	18,77	18,77
	y	559,48	552,02	533,85	539,22	549,42	553,12
5	*x*	17,86	19,29	17,58	18,83	17,75	17,49
	y	537,66	535,68	541,28	546,18	550,84	551,58
6	*x*	18,42	18,15	18,97	17,61	17,91	18,96
	y	557,44	535,81	552,21	540,84	536,83	548,6
7	*x*	18,33	17,2	17,39	18,97	17,94	18,59
	y	555,61	544,63	549,14	558,76	554,71	549,29
8	*x*	17,57	19,31	18,4	17,44	17,58	17,12
	y	557,69	539,13	559,75	557,16	536,89	537,44
9	*x*	19,3	18,84	17,92	17,87	17,98	18,13
	y	530,91	559,49	550,67	548,35	536,02	536,65
10	*x*	17,11	18,17	19,04	18,85	18,71	17,52
	y	532,44	548,24	549,15	534,32	530,48	543,26

Найти описательные статистики экспериментальных данных средствами среды программирования R и Excel.

2. Найти описательные статистики экспериментальных данных средствами среды программирования R и Excel зарплаты (р.) и возраста сотрудника гостиницы по следующим данным

Вариант	Данные						
	Возраст	20	50	45	40	25	30
1	Зарплата	800	2500	2500	2000	1200	1800
2	Зарплата	900	2600	2500	2200	1300	1900
3	Зарплата	700	2700	2500	2300	1100	1700
4	Зарплата	1000	2400	2300	1900	1200	1500
5	Зарплата	900	2400	2400	2000	1400	1700
6	Зарплата	1000	2500	2300	2100	1200	1900
7	Зарплата	700	2500	2400	1900	1300	1600
8	Зарплата	900	2300	2300	1900	1200	1800
9	Зарплата	800	2600	2600	2100	1300	1700
10	Зарплата	800	2600	2500	2200	1200	1800

2 ЛАБОРАТОРНАЯ РАБОТА № 2. Графические возможности среды программирования R для визуализация результатов описательной статистики

Цель работы: Изучение визуализации результатов описательной статистики экспериментальных данных с использованием возможности среды программирования R.

Основные теоретические положения

Для выполнения данной работы понадобится подключить следующие пакеты:

```
library("psych")# описательные статистики
library("lmtest") # тестирование гипотез в линейных моделях library("ggplot2")# графики
library("dplyr") # манипуляции с данными
library("MASS") # подгонка распределений
```

Получим, например, описание набора данных по автомобилям `cars` командой:

```
help(cars)
```

Результат выполнения команды (в правом нижнем углу на вкладке `Help`) показан на рисунке 5.

В этом наборе данных 50 наблюдений и две переменных (ско- рость, миль/час и длина тормозного пути в футах).

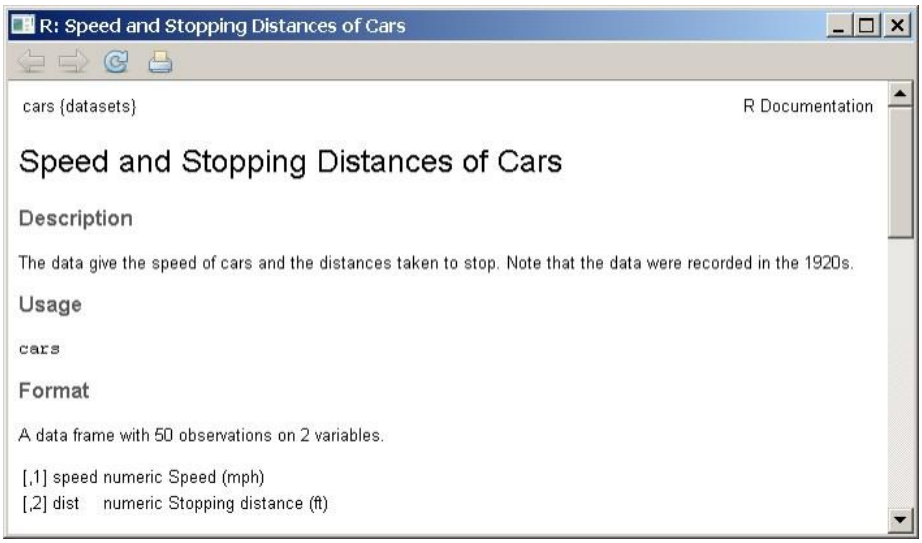


Рис. 5. Справка по набору данных `cars`

Поместим в переменную d встроенный в R набор данных по автомобилям:

```
d <- cars # этот набор данных находится в базовом пакете datasets
```

Теперь `d` имеет тип данных `data.frame` (набор данных), в чем можно удостовериться, посмотрев в правом верхнем углу окна таблицу среды `Environment` (рис. 6).

Для этого должен быть выбран режим `Grid`. С помощью кнопки ▦ можно просмотреть содержимое набора данных в виде таблицы.

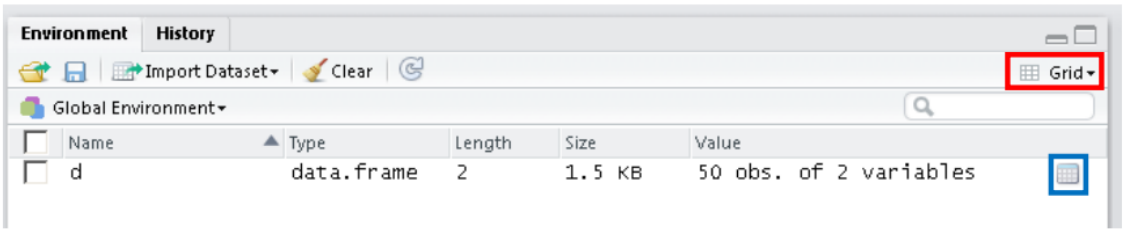


Рис. 6. Описание набора данных `d` в таблице среды `Environment`

Следующей командой можно посмотреть на этот набор данных, в результате чего будут перечислены все переменные и типы данных:

```
glimpse(d) # функция из пакета dplyr
```

Результат выполнения команды появится в консоли:

```
> d <- cars
> glimpse(d)
Observations: 50
Variables: 2
$ speed (dbl) 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13,...
$ dist  (dbl) 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 28,...
```

Переменные `speed` и `dist` имеют тип данных `dbl` (double) и содержат по 50 наблюдений. Для других типов данных используются следующие сокращения: `chr` (character/string), `int` (integer), `fctr` (factor), `tims` (time), `lgl` (logical).

Посмотрим на первые шесть наблюдений набора данных `d`:

```
> head(d) # функция из базового пакета utils
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

и последние шесть наблюдений:

```
> tail(d) # функция из базового пакета utils
   speed dist
45    23   54
46    24   70
47    24   92
48    24   93
49    24  120
50    25   85
```

Получим таблицу с описательными статистиками: среднее, мода, медиана, стандартное отклонение, минимум/максимум, асимметрия, эксцесс и т. д.:

```
> describe(d) # функция из пакета psych
      vars  n  mean    sd median trimmed   mad min max range  skew kurtosis
speed    1 50 15.40  5.29     15   15.47  5.93   4  25    21 -0.11    -0.67
dist     2 50 42.98 25.77     36   40.88 23.72   2 120   118  0.76     0.12
        se
speed 0.75
dist  3.64
```

Построим гистограмму абсолютных частот для переменной `dist` (длины тормозного пути).

Воспользуемся функцией `qplot`, задав источник данных `d` (аргумент `data`), переменную для построения графика (`dist`), подпишем оси (параметры функции `xlab` и `ylab`) и название графика (параметр `main`):

```
# функция из пакета ggplot2
qplot(data=d, dist, xlab="Длина тормозного пути (футы)", ylab="Число автомобилей",main="Данные по автомобилям 1920х")
```

Результат выполнения данной функции показан на рисунке 7.

Рис. 7. Гистограмма абсолютных частот для переменной `dist`

Можно построить также гистограмму плотности распределения (рис. 8):

```
# функция из базового пакета graphics
hist(d$dist, probability = TRUE, col="grey")
```

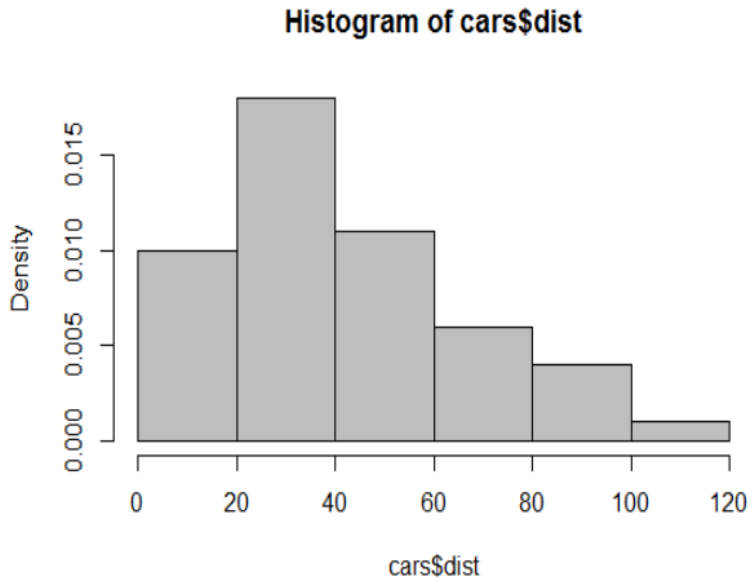


Рис. 8. Гистограмма плотности распределения для переменной `dist`

Подгоним плотность распределения Вейбулла, поместив результат (оценки параметров распределения) в переменную `fit`:

```
fit <- fitdistr(d$dist, "weibull") # функция из пакета MASS
```

Переменная `fit` теперь представляет собой список (`List`) из 5 элементов, что можно увидеть в `Environment` (рис. 9).

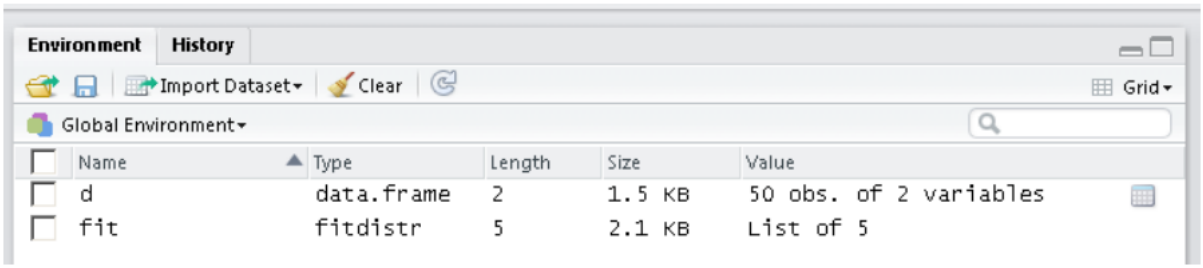


Рис. 9. Описание переменной `fit` в таблице среды `Environment`

Доступ к элементам списка можно получить через значок доллара `$`.

Оценки двух параметров распределения Вейбулла были рассчитаны методом максимального правдоподобия.

Просмотрим их, обратившись к элементу списка `fit`:

```
> fit$estimate
   shape    scale
 1.72371 48.15234
```

Покажем на том же графике теоретическую плотность распределения Вейбулла (рис. 10):

```
xvals <- seq(0, 120, .20)   # значения по оси абсцисс от 0 до 120 с шагом 0,2
lines(xvals, dweibull(xvals, shape=fit$estimate[1],
scale=fit$estimate[2]))
```

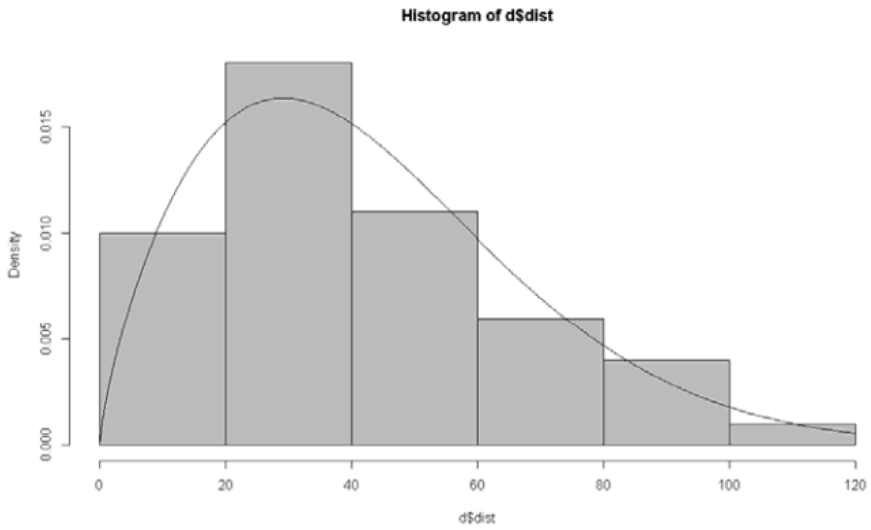


Рис. 10. Гистограмма распределения переменной `dist`

Первый аргумент функции `lines` — это значения по оси абсцисс, на основе которых будет построен график.

Пример

1 Имеются данные о цене на нефть x (ден. ед.) и индексе акций нефтяных компаний y (усл. ед.):

Вариант		Данные					
1	x	17,28	17,05	18,30	18,80	19,20	18,50
	y	537	534	550	555	560	552

Построить зависимость индекса акций нефтяных компаний от цены на нефть.

Решение

```
oil_values <- c(17.28, 17.05, 18.3, 18.8, 19.2, 18.5)
// c(1, 2, .., n) - создаёт вектор с указанными значениями

action_values <- c(537, 534, 550, 555, 560, 552)

vector <- c(oil_values, action_values)
// Создаётся новый вектор, состоящий из двух векторов (они склеиваются)

m <- matrix(ncol = 2, nrow = 6, data = vector)
```

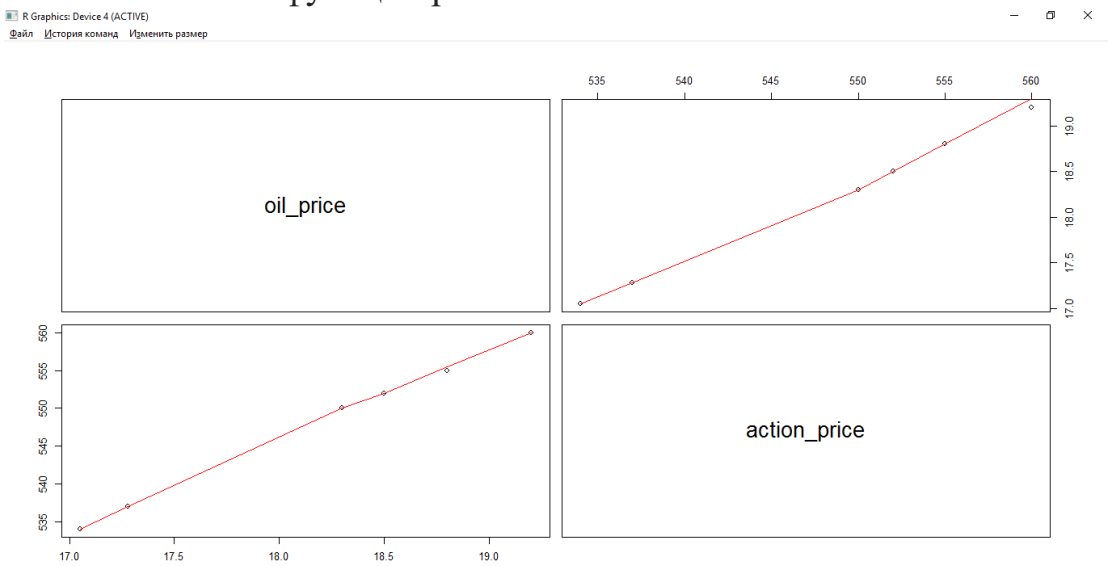
```
// Создаётся матрица, состоящая из 2 столбцов и 6 строк, ячейки
которой заполнены значениями вектора. Матрица заполняется сверху
вниз, слева направо.

dimnames(m) = list(  c('T1', 'T2', 'T3', 'T4', 'T5', 'T6'),
c('oil_price', 'action_price'))
// Подписываются столбцы и строки. T1, T2 ... - какие-то моменты
времени. oil_price, action_price - столбцы с ценами нефти и акций

x11()
// Создание отдельного окна, в котором будет выводиться графиче-
ская информация.

pairs(m, panel = panel.smooth)
```

Результат выполнения функции pairs.



У таблицы два столбца. График на пересечении oil_price и action_price (слева внизу) показывает зависимость цены акций от цены на нефть.

Задания для лабораторной работы

1. Имеются данные о цене на нефть x (ден. ед.) и индексе акций нефтяных компаний y (усл. ед.).

Вариант №	Данные						
1	x	18,26	18,06	19,40	19,70	20,10	19,60
	y	567	564	570	575	580	572
2	x	18,71	19,44	18,65	17,18	17,77	17,39
	y	556,3	555,94	544,9	540,72	541,07	559,47
3	x	18,99	17,71	18,82	17,42	18,4	18,15
	y	549,66	555,85	541,18	544,68	537,08	538,97
4	x	19,11	18,72	18,56	18,99	18,77	18,77
	y	559,48	552,02	533,85	539,22	549,42	553,12
5	x	17,86	19,29	17,58	18,83	17,75	17,49
	y	537,66	535,68	541,28	546,18	550,84	551,58
6	x	18,42	18,15	18,97	17,61	17,91	18,96
	y	557,44	535,81	552,21	540,84	536,83	548,6

7	*x*	18,33	17,2	17,39	18,97	17,94	18,59
	y	555,61	544,63	549,14	558,76	554,71	549,29
8	*x*	17,57	19,31	18,4	17,44	17,58	17,12
	y	557,69	539,13	559,75	557,16	536,89	537,44
9	*x*	19,3	18,84	17,92	17,87	17,98	18,13
	y	530,91	559,49	550,67	548,35	536,02	536,65
10	*x*	17,11	18,17	19,04	18,85	18,71	17,52
	y	532,44	548,24	549,15	534,32	530,48	543,26

Построить зависимость индекса акций нефтяных компаний от цены на нефть средствами Excel и R.

2. Построить зависимость зарплаты (р.) от возраста сотрудника гостиницы по следующим данным (средствами Excel и R).

Вариант	Данные						
	Возраст	20	50	45	40	25	30
1	Зарплата	800	2500	2500	2000	1200	1800
2	Зарплата	900	2600	2500	2200	1300	1900
3	Зарплата	700	2700	2500	2300	1100	1700
4	Зарплата	1000	2400	2300	1900	1200	1500
5	Зарплата	900	2400	2400	2000	1400	1700
6	Зарплата	1000	2500	2300	2100	1200	1900
7	Зарплата	700	2500	2400	1900	1300	1600
8	Зарплата	900	2300	2300	1900	1200	1800
9	Зарплата	800	2600	2600	2100	1300	1700
10	Зарплата	800	2600	2500	2200	1200	1800

3 ЛАБОРАТОРНАЯ РАБОТА № 3. Принятие статистических решений с помощью параметрических и непараметрических критериев средствами среды программирования R.

Цель работы: Изучение параметрических и непараметрических критериев средствами среды программирования R.

Основные теоретические положения

Статистическая гипотеза – это предположение о виде или отдельных параметрах распределения вероятностей, которое подлежит проверке на имеющихся данных.

Проверка статистических гипотез – это процесс формирования решения о возможности принять или отвергнуть утверждение (гипотезу), основанный на информации, полученной из анализа выборки. Методы проверки гипотез называются критериями.

В большинстве случаев рассматривают так называемую нулевую гипотезу (нуль-гипотезу H_0), состоящую в том, что все события произошли случайно,

естественным образом. Альтернативная гипотеза H_1 состоит в том, что события случайным образом произойти не могли, и имело место воздействие некоторого фактора.

Обычно нулевая гипотеза формулируется таким образом, чтобы на основании эксперимента или наблюдений ее можно было отвергнуть с заранее заданной вероятностью ошибки. Эта заранее заданная вероятность ошибки называется **уровнем значимости**.

Уровень значимости – максимальное значение вероятности появления события, при котором событие считается практически невозможным. В статистике наибольшее распространение получил уровень значимости $\alpha = 0{,}05$. Поэтому, если вероятность, с которой интересующее событие может произойти случайным образом $p < 0{,}05$, то принято считать это событие маловероятным, и если оно все же произошло, то это не было случайным. В наиболее ответственных случаях, когда требуется особая уверенность в достоверности полученных результатов, надежности выводов, уровень значимости принимают равным $\alpha = 0{,}01$ или даже $\alpha = 0{,}001$.

Величину $P = 1 - \alpha$ называют доверительной вероятностью (уровнем надежности), то есть вероятностью, признанной достаточной для того, чтобы уверенно судить о принятом статистическом решении. Соответственно, в качестве доверительных вероятностей выбирают значения 0,95, 0,99 или 0,999.

Интервал, в котором с заданной доверительной вероятностью $P = 1 - \alpha$ находится оцениваемый параметр, называется **доверительным интервалом**. В соответствии с доверительными вероятностями на практике используются 95-, 99-, 99,9-процентные доверительные интервалы. Граничные точки доверительного интервала называют доверительными пределами.

В MS Excel для более точного вычисления границ доверительного интервала при числе элементов в выборке $n < 30$ можно воспользоваться функцией ДОВЕРИТ или процедурой Описательная статистика.

Функция ДОВЕРИТ(альфа; станд откл; размер) определяет полуширину доверительного интервала и содержит следующие параметры:

альфа – уровень значимости, используемый для вычисления доверительной вероятности. Доверительная вероятность равняется 100*(1 – альфа) процентам, или, другими словами, альфа равное 0,05 означает 95-процентный уровень доверительной вероятности;

станд_откл – стандартное отклонение генеральной совокупности для интервала данных, предполагается известным;

размер – это размер выборки.

Пример. Найти границы 95-процентного доверительного интервала для среднего значения, если у 25 телефонных аккумуляторов среднее время разряда в режиме ожидания составило 140 часов, а стандартное отклонение – 2,5 часа.

Решение.

1. Откройте новую рабочую таблицу. Установите табличный курсор в ячейку A1.

2. Для определения границ доверительного интервала необходимо на па-

нели инструментов Формулы нажать кнопку Вставка функции. В появившемся диалоговом окне Мастера функций выберите категорию Статистические и функцию ДОВЕРИТ.НОРМ, после чего нажмите кнопку ОК.

3. В рабочие поля появившегося диалогового окна функции ДОВЕРИТ.НОРМ с клавиатуры введите условия задачи: Альфа – 0,05; Станд_откл – 2,5; Размер – 25 (рисунок 5). Нажмите кнопку ОК.

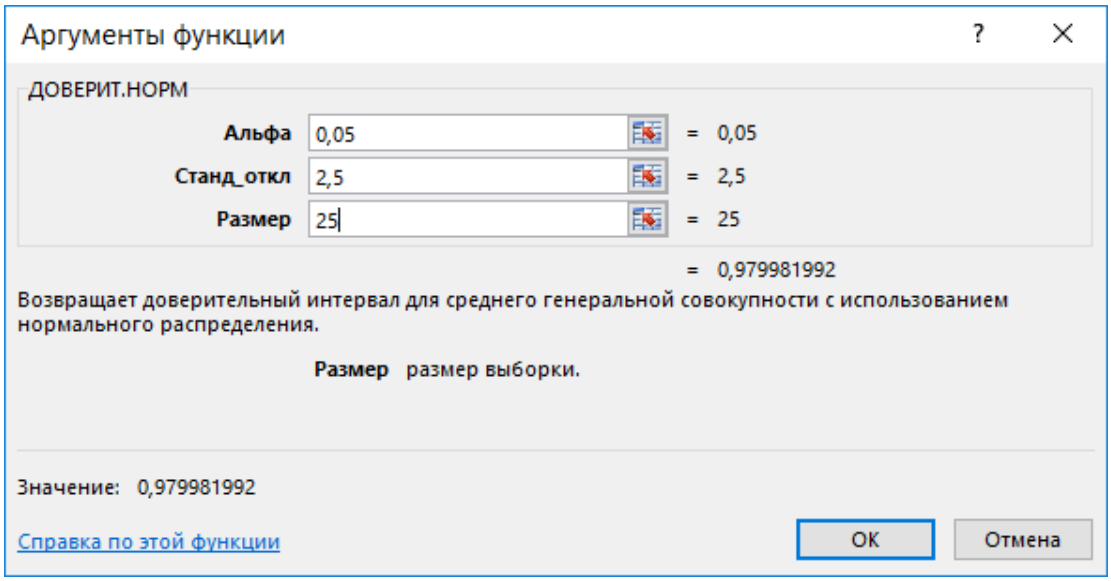


Рисунок 5 – Пример заполнения диалогового окна ДОВЕРИТ.НОРМ

4. В ячейке А1 появится полуширина 95-прцснтного доверительного интервала для среднего значения выборки – 0,979981. Другими словами, с 95-процентным уровнем надежности можно утверждать, что средняя продолжительность разряда аккумулятора составляет 140 ± 0,979981 часа или от 139,02 до 140,98 часа.

Критерий $\chi 2$ используется для анализа таблиц сопряженности признаков и сравнения законов распределения непрерывных случайных величин. Анализируются номинальные или приведенные к номинальной шкале данные, представленные в виде таблицы сопряженности признаков. Для непрерывных случайных величин используется принадлежность значений заданным интервалам, выбираемых таким образом, чтобы в каждом из них было не менее 5-7 значений (интервалы с меньшим числом значений объединяются). Простейшим выбором является равный шаг интервалов, равный $\lambda = \frac{x_{\max} - x_{\min}}{k}$, $k = 1 + 3.32 \cdot \lg(n)$ или $k = 5 \cdot \lg(n)$.

Вычисление критериальной статистики производится по формуле:

$$\chi^2 = \sum_i \frac{(n_i - n_i^{'})^2}{n_i^{'}}$$

где n_i- эмпирические частоты, $n_i^{'}$ - теоретические частоты попадания элементов выборки в группы (заданные интервалы).

Число степеней свободы находят по формуле:

$$f = k - 1 - r$$

где k-число групп выборки, r-число параметров предполагаемого распределения, которые оценены по данным выборки.

Если предполагаемое распределение – нормальное, то по выборке оценивают два параметра (математическое ожидание и дисперсию), поэтому r=2 и $f = k - 3$.Одной из функций, осуществляющей проверку данного критерия в R является **`chisq.test()`**.

Описание функции

`chisq.test (x, y = NULL, p = rep(1/length (x), length (x)))`

Параметры

`x`	вектор или матрица.
`y`	вектор; игнорируемый, если x - матрица.
`p`	вектор теоретических вероятностей той же длины, что x.

Примечание

Если `x` – матрица с одной строкой или столбцом, или если `x` – вектор, и `y` не дан, `x` – одномерная таблица сопряженности признаков. В этом случае, проверенная гипотеза - равняются ли вероятности совокупности тем, что в `p`, или все равны, если `p` не дается. Если `x` - матрица с двумя строками (или столбцами), содержащими неотрицательные целые числа, то она рассматривается как таблица сопряженности признаков. Если `x` и `y` – два

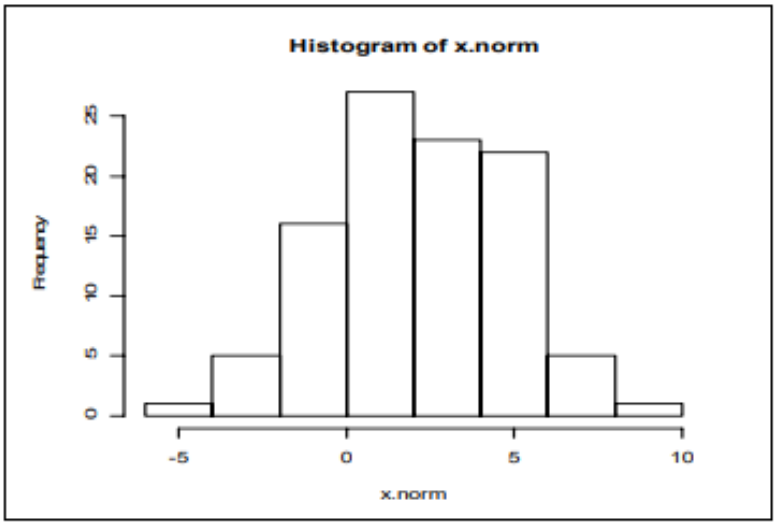
вектора, содержащих факторы (номинальные или ординальные значения), то по ним строится таблица сопряженности.

Критические значения (квантили) находятся с использованием функции **qchisq(p,df)** или по таблице χ2-распределения [2,стр.329].

Пример

Сгенерируем случайную выборку из нормального распределения, и проверим ее нормальность.

```
N<-100  # объем выборки
x.norm<-rnorm(N,mean=2,sd=2.5) # задаем среднее и СКО
```



Вычисляем квантили выборки с шагом 10% (по 10 элементов в интервале)

```
> x.norm.q <- quantile(x.norm,probs=seq(0,1,0.1))
> round(x.norm.q,2)
   0%   10%   20%   30%   40%   50%   60%   70%   80%
90%  100%
 -4.12 -1.51 -0.14  0.59  1.52  2.15  2.70  3.89  4.51
5.22  8.15
> summary(x.norm)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.4490  0.6675  2.0330  2.0670  3.1050  6.5830
```

Выбираем интервалы:

```
> k<-6    # число интервалов
> x.q <- c(-10, -1.0, 0.5, 2.0, 3.5, 5.0,12.0)
```

Вычисляем фактические частоты

```
> x.norm.hist<-hist(x.norm,breaks=x.q,plot=FALSE)
> x.norm.hist$counts
12 15 22 18 18 15
```

Вычисляем (по выборке) теоретические вероятности для каждого интервала

```
> x.q[1]<-(-Inf);x.q[k+1]<-(+Inf)#«раздвигаем» границы
до бесконечности
```

```
  > x.norm.p.theor<-
pnorm(x.q,mean=mean(x.norm),sd=sd(x.norm))
  > x.norm.p.theor<-(x.norm.p.theor[2:(k+1)]-
x.norm.p.theor[1:k])
  > round(x.norm.p.theor,2)
  0.12 0.15 0.21 0.22 0.16 0.14
```

Сравниваем фактические и теоретические частоты

```
  > chisq.test(x.norm.hist$counts,p=x.norm.p.theor)
          Chi-squared test for given probabilities
  data:  x.norm.hist$counts
  X-squared = 0.9691, df = 5, p-value = 0.965
```

Поскольку для проверки нулевой гипотезы H_0 о нормальности распределения генеральной совокупности в нашем случае используется правосторонний критерий, а уровень значимости (*p*-value) равен 0.965 (96.5%), то нужно допустить разрешить вероятность ошибки, равную 96.5%, чтобы считать выборку не принадлежащей нормальному распределению. Следовательно, гипотеза о нормальности принимается.

Приведем пример проверки случайности сопряжения признаков. Пусть первый признак является полом (закодированным символами «М» и «Ж»), а второй – наличием близорукости (закодированной числами 0, 1).

```
  > x<-
c("М","Ж","Ж","М","Ж","Ж","М","Ж","М","Ж","Ж","М","М")
  > y<-c( 0,  0,  1,  0,  0,  1,  0,  0,  1,  1,  0,  1,
0)
  > tbl<-table(x,y)
  > tbl
     y
  x  0 1
  Ж  4 3
  М  4 2
```

Оцениваем по выборке маргинальные (безусловные) вероятности для x и y:

```
  > p.x<-c(sum(tbl[1,])/sum(tbl),sum(tbl[2,])/sum(tbl))
  > p.y<-c(sum(tbl[,1])/sum(tbl),sum(tbl[,2])/sum(tbl))
```

или, что тоже самое (годится для таблиц любой размерности),

```
  > p.x<-apply(tbl,1,sum)/sum(tbl) # 1 - суммируем по
строкам
  > p.y<-apply(tbl,2,sum)/sum(tbl) # 2 - суммируем по
столбцам
```

Если сопряженности признаков нет (гипотеза H_0), то наблюдаемые относительные частоты должны совпадать с произведениями маргинальных вероятностей:

```
  >p.theor<-c(p.x[1]*p.y[1],p.x[2]*p.y[1],
p.x[1]*p.y[2],p.x[2]*p.y[2])
```

или, что тоже самое (годится для таблиц любой размерности),

```
> p.theor<- p.x %*% t(p.y)
> chisq.test(as.vector(t),p=p.theor)
        Chi-squared test for given probabilities

data:  as.vector(t)
X-squared = 0.1238, df = 3, p-value = 0.9888
```

Поскольку *p*-value близко к единице, то нулевая гипотеза принимается. Отметим, что поскольку в таблице сопряженности tbl есть ячейки с числами, меньшими 5 (в нашем случае – все ячейки такие), то аппроксимация хи-квадрат может быть неправильной, а полученный результат - неверным.

Задания для лабораторной работы

1 Определить, лежит ли значение X внутри границ 95-процентного доверительного интервала выборки:

Вариант №	X	Данные
1	19	2, 3, 5, 7, 4, 9, 6, 4, 9, 10, 4, 7, 19.
2	13	9, 4, 9, 10, 4, 7, 8, 4, 4, 5, 10, 6, 13.
3	15	7, 1, 4, 10, 9, 7, 3, 3, 10, 3, 10, 10, 15.
4	14	9, 4, 5, 1, 7, 2, 8, 5, 6, 2, 1, 5, 14.
5	15	9, 3, 3, 1, 9, 5, 6, 5, 6, 8, 9, 7, 15.
6	11	3, 8, 2, 1, 6, 9, 8, 6, 5, 10, 1, 1, 11.
7	18	2, 8, 7, 3, 7, 6, 1, 5, 10, 10, 2, 6, 18.
8	16	5, 3, 4, 6, 8, 6, 4, 2, 5, 9, 8, 10, 16.
9	13	1, 8, 10, 7, 5, 5, 1, 9, 3, 2, 9, 2, 13.
10	10	8, 4, 9, 1, 8, 4, 7, 4, 6, 6, 6, 4, 10.

2 Определите с уровнем значимости a = 0,05 максимальное отклонение среднего значения генеральной совокупности от среднего выборки:

Вариант №	Данные
1	3, 4, 4, 2, 5, 3, 4, 3, 5, 4, 3, 5, 6.
2	6, 4, 4, 6, 2, 4, 6, 3, 6, 4, 2, 5, 2.
3	5, 4, 4, 5, 3, 3, 6, 5, 6, 2, 2, 2, 5.
4	4, 2, 4, 4, 2, 2, 6, 4, 2, 6, 4, 5, 6.
5	3, 4, 6, 6, 3, 2, 6, 6, 5, 5, 5, 6, 6.
6	3, 2, 5, 2, 4, 4, 5, 4, 6, 5, 6, 3, 3.
7	6, 5, 5, 4, 4, 2, 2, 4, 6, 3, 5, 2, 5.
8	6, 3, 5, 5, 4, 5, 5, 3, 6, 6, 5, 2, 4.
9	2, 6, 4, 4, 5, 2, 4, 6, 2, 2, 2, 3, 3.
10	5, 6, 3, 4, 6, 2, 6, 5, 4, 4, 5, 2, 3.

3 Найти соответствие экспериментальных данных нормальному закону распределения для следующей выборки весов детей (кг):

Вариант №	Данные
1	21, 21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 27, 27.
2	30, 30, 30, 30, 30, 30, 30, 30, 30, 31, 31, 31, 31, 31, 31, 31, 32, 32, 32, 32, 32, 32, 33, 33, 33, 33, 33, 33, 33, 33, 33, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 35, 35, 35, 35, 35, 35, 35, 35, 35, 36, 36, 36, 36, 36, 36, 36, 36, 36.
3	15, 15, 15, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 17, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22, 22.
4	10, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12, 12, 12, 13, 13, 13, 13, 13, 13, 13, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16.
5	23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 27, 27, 27, 27, 27, 27, 27, 27, 27, 28, 28, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29, 29, 29.
6	20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26, 26, 26.
7	5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12, 12.
8	21, 21, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27.
9	17, 17, 17, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 24, 24.
10	26, 26, 26, 26, 26, 26, 26, 27, 27, 27, 27, 27, 27, 27, 27, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29, 29, 30, 30, 30, 30, 30, 30, 30, 30, 30, 31, 31, 31, 31, 31, 31, 31, 31, 32, 32, 32, 32, 32, 32, 32, 32, 33, 33.

4 Даны результаты бега на дистанцию 100 м в секундах в двух группах студентов. Студенты первой группы в течение года посещали факультативные занятия по физкультуре. Определить, достоверны ли отличия по результатам бега в этих группах.

Вариант №	Данные	
	Посещавшие факультатив	Не посещавшие
1	12,6, 12,3, 11,9, 12,2, 13,0, 12,4.	12,8, 13,2, 13,0, 12,9, 13,5, 13,1.
2	12,9, 12,4, 12,3, 12,2, 12,1, 12,1.	13,4, 12,8, 12,1, 13,3, 12,2, 12,9.
3	12,4, 12,4, 12,3, 12,1, 12,3, 12,2.	12,3, 13,0, 12,1, 12,3, 12,0, 12,2.
4	12,7, 12,5, 12,6, 12,8, 13,0, 12,4.	12,9, 13,4, 13,2, 13,2, 12,2, 13,3.
5	12,6, 12,4, 12,5, 12,3, 12,4, 12,5.	12,9, 12,9, 13,1, 13,3, 13,2, 13,3.
6	12,1, 12,5, 12,9, 12,7, 12,4, 12,1.	12,7, 12,8, 13,4, 12,3, 13,1, 12,3.
7	12,1, 12,1, 12,6, 12,7, 12,7, 12,8.	13,4, 12,6, 13,1, 13,4, 13,2, 12,2.
8	12,6, 12,1, 12,8, 12,9, 12,6, 12,7.	13,1, 13,2, 13,5, 12,3, 13,1, 13,4.
9	12,5, 12,5, 13,0, 12,6, 12,6, 12,7.	12,9, 12,5, 12,7, 13,3, 13,5, 13,0.
10	12,0, 12,2, 12,2, 12,7, 12,4, 12,7.	13,3, 12,4, 12,1, 12,5, 12,1, 13,2.

4 ЛАБОРАТОРНАЯ РАБОТА № 4. Проведение корреляционного анализа средствами среды программирования R

Цель работы: Изучение способов оценки степени взаимосвязи между наблюдаемыми переменными с помощью табличного процессора Excel и среды программирования R.

Основные теоретические положения

Важным разделом статистического анализа является корреляционный анализ, служащий для выявления взаимосвязей между выборками.

Выявление взаимосвязей. Одна из наиболее распространенных задач статистического исследования состоит в изучении связи между некоторыми наблюдениями переменных. Знание взаимозависимостей отдельных признаков дает возможность решать одну из кардинальных задач любого научного исследования: возможность предвидеть, прогнозировать развитие ситуации при изменении конкретных характеристик объекта исследования.

Обычно взаимосвязь между выборками носит не функциональный, а вероятностный (или стохастический) характер. В этом случае нет строгой, однозначной зависимости между величинами. При изучении стохастических зависимостей различают корреляцию и регрессию.

Регрессионный анализ устанавливает формы зависимости между случайной величиной и значениями одной или нескольких переменных величин.

Корреляционный анализ состоит в определении степени связи между двумя случайными величинами X и Y. В качестве меры такой связи используется *коэффициент корреляции.* Он оценивается по выборке объема n связанных пар наблюдений (x_i, y_i) из совместной генеральной совокупности X и Y. Существует несколько типов коэффициентов корреляции, применение которых зависит от предположений о совместном распределении величин X и Y.

Для оценки степени взаимосвязи наибольшее распространение получил коэффициент линейной корреляции (Пирсона), предполагающий нормальный закон распределения наблюдений.

Коэффициент корреляции (R, r) – параметр, характеризующий степень линейной взаимосвязи между двумя выборками. Коэффициент корреляции изменяется от -1 (строгая обратная линейная зависимость) до 1 (строгая прямая пропорциональная зависимость). При значении коэффициента равном 0 линейной зависимости между двумя выборками нет. Здесь под прямой зависимостью понимают зависимость, при которой увеличение или уменьшение значения одного признака ведет, соответственно, к увеличению или уменьшению второго.

Выборочный коэффициент линейной корреляции между двумя случайными величинами X и Y рассчитывается по формуле

$$r = \frac{\sum(x - M_x)(y - M_y)}{\sqrt{\sum(x - M_x)^2 (y - M_y)^2}}$$

Коэффициент корреляции является безразмерной величиной, и его значение не зависит от единиц измерения случайных величин X и Y.

На практике коэффициент корреляции принимает некоторые промежуточные

значения между 1 и -1. Для оценки степени взаимосвязи можно руководствоваться следующими эмпирическими правилами. Если коэффициент корреляции r по абсолютной величине (без учета знака) больше, чем 0,95, то принято считать, что между параметрами существует практически линейная зависимость (прямая – при положительном r и обратная – при отрицательном r). Если коэффициент корреляции $|r|$ лежит в диапазоне от 0,8 до 0,95, говорят о сильной степени линейной связи между параметрами. Если $0,6 < |r| < 0,8$, говорят о наличии линейной связи между параметрами. При $|r| < 0,4$ обычно считают, что линейную взаимосвязь между параметрами выявить не удалось.

В MS Excel для вычисления парных коэффициентов линейной корреляции используется специальная функция КОРРЕЛ. Функция имеет следующий синтаксис:

КОРРЕЛ(массив1: массив2),

где Массив1 – это диапазон ячеек первой случайной величины;

Массив2 – это второй интервал ячеек со значениями второй случайной величины.

Пример. Имеются результаты семимесячных наблюдений реализации путевок двух туристских маршрутов тура А и тура В (таблица 2).

Таблица 2 – Результаты семимесячных наблюдений

Тур А	120	121	105	92	112	91	80
Тур В	20	19	17	16	18	16	15

Необходимо определить, имеется ли взаимосвязь между количеством продаж путевок обоих маршрутов.

Решение. Для выявления степени взаимосвязи прежде всего необходимо ввести данные в рабочую таблицу. Затем вычисляется значение коэффициента корреляции между выборками. Для этого установите курсор в свободную ячейку (например А9). Вызовите функцию КОРРЕЛ. Введите в иоле Массив1 диапазон данных А2:А8. В поле Массив2 введите диапазон данных В2:В8. Нажмите кнопку ОК. В ячейке А9 появится значение коэффициента корреляции – 0,969123. Значение коэффициента корреляции больше чем 0,95. Значит, можно говорить о том, что в течение периода наблюдения имелась высокая степень прямой линейной взаимосвязи между количествами проданных путевок обоих маршрутов (r = 0,969123).

Корреляционная матрица. При большом числе наблюдений, когда коэффициенты корреляции необходимо последовательно вычислять из нескольких рядов числовых данных, для удобства получаемые коэффициенты сводят в таблицы, называемые корреляционными матрицами.

Корреляционная матрица – это квадратная (или прямоугольная) таблица, в которой на пересечении соответствующих строки и столбца находится коэффициент корреляции между соответствующими параметрами.

В MS Excel для вычисления корреляционных матриц используется процедура Корреляция. Процедура позволяет получить корреляционную матрицу, содержащую коэффициенты корреляции между различными параметрами.

Для реализации процедуры необходимо:

- выполнить команду Сервис – Анализ данных;
- в появившемся списке Инструменты анализа выбрать строку Корреляция и нажать кнопку ОК;
- в появившемся диалоговом окне указать Входной интервал, то есть ввести ссыл-

ку на ячейки, содержащие анализируемые данные. Входной интервал должен содержать не менее двух столбцов:

- в разделе Группировка переключатель установить в соответствии с введенными данными (например, по столбцам);
- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа. Для этого следует установить флажок Выходной интервал, далее навести указатель мыши на правое поле ввода Выходной интервал и щелкнуть левой кнопкой мыши, затем указатель мыши навести на левую верхнюю ячейку выходного диапазона и щелкнуть левой кнопкой мыши. Размер выходного диапазона будет определен автоматически и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные (рисунок 7);
- нажать кнопку ОК.

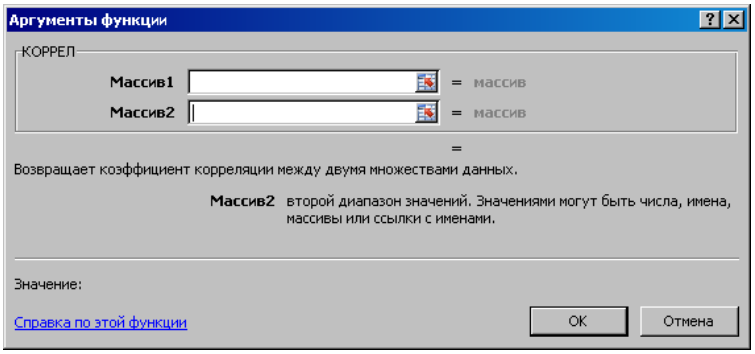


Рисунок 7 – Пример установки параметров корреляционного анализа

Коэффициент корреляции Пирсона

Выборочный коэффициент корреляции Пирсона определяется равенством:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot S_x \cdot S_y}$$

x_i, y_i - варианты зависимой (парной) выборки $\langle x, y \rangle$, т.е. выборочные значения признаков X и Y, выбираемых парами, n -объем выборки, S_x и S_y - выборочные средние квадратические отклонения, $\bar{x}$ и $\bar{y}$ - выборочные средние.

Известно, что если величины Y и X независимы, то коэффициент корреляции генеральных совокупностей $\rho_{xy}=0$; если $\rho_{xy}=\pm 1$, то Y и X связаны линейной функциональной зависимостью, т.е. коэффициент корреляции ρ_{xy} (а значит и выборочный коэффициент корреляции r_{xy}, являющийся оценкой для ρ_{xy}) измеряют силу линейной связи между Y и X. В системе R для вычисления выборочного коэффициента корреляции используется функция cor().

Если основная гипотеза H_0 говорит о независимости Y и X (т.е. коэфициент корреляции равен нулю), то в качестве критериальной статистики используется статистика

$$t_p = \sqrt{n-2} \cdot \frac{r_{xy}}{\sqrt{1-r_{xy}^2}},$$

имеющая распределение Стьюдента с $n-2$ степенями свободы. Если объем выборки большой ($n \sim 100$), то можно считать, что статистика имеет нормальное распределение.

Проверка гипотезы о значимости показателя ранговой корреляции осуществляется функцией **cor.test()**

Описание функции

cor(x, y, method = c("pearson", "kendall", "spearman"))

Параметры

X	Вектор, матрица или data.frame
Y	Второй вектор (или NULL, если первый аргумент – матрица или фрейм данных)
method	Вычисляемый коэфициент корреляции (по умолчанию – pearson)

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.8,6.4)
> cor(x, y)
0.4668
```

Коэффициент ранговой корреляции

В качестве критериев оценки назависимости могут применяться и другие коэффициенты корреляции, например показатель ранговой корреляции Спирмена, позволяющий оценить нелинейную, но монотонную зависимость: в этом случае вычисляется кореляция не самих значений, а их рангов (порядковых номеров при упорядочении). Другим ранговым критерием является τ-критерий Кендалла.

Проверка по нескольким критериям может быть использована для приблизительной оценки оценки вида зависимости: если ранговая корреляция большая (статистически значимая), а линейная – маленькая (статистически не значимая), то зависимость нелинейная; если обе корреляции большие, то зависимость линейная; если обе корреляции маленькие, что либо зависимости нет, либо она немонотонная.

Если основная гипотеза гласит, что коэфициент корреляции равен не нулю, а некоторому отличному от нуля числу, то в качестве критериальной статистики используется *z*-преобразование Фишера:

$$z(r_{xy}) = \frac{1}{2}\ln\frac{1+r_{xy}}{1-r_{xy}}$$

Эта величина распределена примерно нормально для всех значений коэффициента корреляции генеральных совокупностей, ее матожидание равно $z(\rho_{xy})$, а дисперсия $1/(n-3)$, где n - объем выборки. Поэтому границы доверительного интервала для $z(\rho_{xy})$ находят с использованием квантилей нормального распределения; получить границы для ρ_{xy} можно обратным преобразованием.

Описание функции

cor.test(x, y,alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), conf.level = 0.95, ...)

Параметры

`x, y`	Числовые вектора x и y одинаковой длины .
`alternative`	Выбирает альтернативную гипотезу одну из "two.sided" (по умолчанию)-двустороняя критическая область, "greater" -правостороняя критическая область или "less"-левостороняя критическая область.
`method`	Выбирает какой коэфициент корреляции используется в тесте. Один из"pearson", "kendall", или "spearman".
`conf.level`	Доверительная вероятность

Примечание

Для проверки нулевой гипотезы H0 о равенстве показателя корреляции нулю необходимо в `alternative` выбрать `"two.sided"`.

Критическое значение находят по таблице критических точек распределения Стьюдента с числом степей свободы $f = n-2$ (в R используется функция вычисления квантилей распределения Стьюдента **qt(p,df)**).

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.8,6.4)
> cor.test(x,y ,alternative = c("two.sided"),
method = c("pearson"))

        Pearson's product-moment correlation

t = 0.9142, df = 3, p-value = 0.428
95 percent confidence interval: -0.7063858  0.9555364
```

```
sample estimates: cor = 0.4667999

> cor.test(x,y,alternative= c("two.sided"),
method=c("spearman"))

        Spearman's rank correlation rho

S = 16, p-value = 0.7833
sample estimates: rho = 0.2
```

Значение

Для обычной линейной корреляции (Пирсона) мы получили выборочное значений 0.4668, значение *t*- статистики 0.9142 при 3 степенях свободы, и *p*-value равное 0.428. Это означает, что отвергнуть нулевую гипотезу можно только при допущении ошибки в 42.8%. 95% доверительный интервал равен (-0.7063858, 0.9555364) и поскольку он содержит ноль, то нулевая гипотеза принимается на 5% уровне значимости.

Для ранговой корреляции Спирмена выборочное значений коэффициента корреляции еще меньше (0.2), а *p*-value еще больше (0.7833). Поэтому и по ранговому критерию мы отвергаем наличие связи между X и Y.

Контрольные задания для лабораторной работы

1 Определить, влияет ли фактор образования на уровень зарплаты сотрудников в гостинице на основании следующих данных:

Вариант №	Данные						
	Образование	Зарплата сотрудника					
1	Высшее	3200	3000	2600	2000	1900	1900
	Среднее спец.	2600	2000	2000	1900	1800	1700
	Среднее	2000	2000	1900	1800	1700	1700
2	Высшее	2500	2700	2000	2100	2900	2600
	Среднее спец.	2000	2400	2100	1900	2100	1800
	Среднее	2300	1900	1800	1700	2000	1900
3	Высшее	2100	3300	2700	2600	2300	2500
	Среднее спец.	2000	2200	1900	2500	2300	2000
	Среднее	2200	1900	1700	2000	2100	1800
4	Высшее	2500	2600	2300	2500	2800	2400
	Среднее спец.	2400	2000	2300	1900	1700	1800
	Среднее	1700	1900	2000	1800	2100	1900
5	Высшее	2600	2300	2700	2500	2200	2400
	Среднее спец.	2300	2000	2200	1900	2400	2300
	Среднее	1900	1700	2100	1800	1800	1900
6	Высшее	2400	2600	2300	2400	2100	2300
	Среднее спец.	2200	1900	2300	2400	2000	2100
	Среднее	1700	1900	2100	2000	1800	1800
7	Высшее	2600	2400	2300	2500	2700	2500
	Среднее спец.	2100	2300	2500	2200	2400	2000
	Среднее	1900	2100	1800	2000	1700	1800
8	Высшее	2700	2900	2500	3200	2600	3000

	Среднее спец.	2400	2200	2700	2300	2500	2000
	Среднее	1900	1600	1800	2000	1900	2200
9	Высшее	2600	2700	3100	2800	2500	2900
	Среднее спец.	2400	2100	2600	2300	2100	2500
	Среднее	2300	2000	2100	1700	1900	1800
10	Высшее	3400	3000	2800	2500	2700	3100
	Среднее спец.	2400	2700	2200	2300	2500	2100
	Среднее	1900	1700	2000	2100	1800	1900

2 Определить, имеется ли взаимосвязь между рождаемостью и смертностью (количество на 1000 человек) в Санкт-Петербурге:

Вариант №	Данные		
	Годы	Рождаемость	Смертность
1	1991	9,3	12,5
	1992	7,4	13,5
	1993	6,6	17,4
	1994	7,1	17,2
	1995	7,0	15,9
	1996	6,6	14,2
2	1991	8,0	16,6
	1992	6,9	16,1
	1993	7,0	13,6
	1994	8,8	14,2
	1995	7,0	17,5
	1996	9,8	11,0
3	1991	7,3	17,8
	1992	9,8	17,6
	1993	8,8	15,3
	1994	7,5	12,4
	1995	8,2	14,1
	1996	6,8	14,7
4	1991	8,1	16,1
	1992	8,4	13,2
	1993	6,9	13,3
	1994	7,0	11,7
	1995	6,7	17,1
	1996	9,4	13,8
5	1991	9,7	16,6
	1992	9,2	14,6
	1993	7,2	12,8
	1994	7,9	15,6
	1995	9,4	15,9
	1996	6,3	12,3
6	1991	10,0	12,6
	1992	8,3	13,0
	1993	8,5	15,7
	1994	7,5	15,8
	1995	6,6	14,7
	1996	6,3	15,2
7	1991	6,0	14,2

	1992	6,8	14,7
	1993	9,1	12,7
	1994	6,1	16,4
	1995	10,0	17,3
	1996	6,4	12,5
8	1991	8,1	12,3
	1992	8,6	15,1
	1993	9,2	13,6
	1994	7,6	15,5
	1995	8,2	15,2
	1996	6,9	15,5
9	1991	7,2	17,3
	1992	6,9	12,9
	1993	7,4	13,4
	1994	8,0	16,8
	1995	9,2	17,0
	1996	6,4	17,7
10	1991	6,7	14,2
	1992	10,0	17,4
	1993	6,6	14,3
	1994	9,3	16,7
	1995	8,3	14,3
	1996	6,7	12,2

3 Определить, имеется ли взаимосвязь между годовым уровнем инфляции (%), ставкой рефинансирования (%) и курсом доллара (р./$), по следующим данным ежегодных наблюдений:

Вариант №	Данные		
	Уровень инфляции	Ставка рефинансирования	Курс $
1	84	85	6,3
	45	55	14
	56	65	20
	34	40	28
	23	28	29
2	80	77	7
	53	50	13
	61	62	21
	38	42	29
	25	26	33
3	79	84	8,2
	47	53	16
	60	67	22
	36	44	29
	27	33	35
4	92	90	7,6
	66	72	12
	78	84	16
	45	47	21
	34	39	29

5	77	78	8,1
	49	52	13
	57	63	17
	38	41	21
	26	29	26
6	84	86	7,9
	53	57	11
	67	70	18
	42	45	22
	30	36	28
7	77	81	8,5
	50	53	12
	63	68	16
	42	47	21
	31	34	26
8	86	90	9,2
	54	62	12
	69	78	18
	42	45	23
	34	37	29
9	80	84	7,6
	52	56	10
	66	73	16
	42	43	20
	31	28	25
10	79	82	9,8
	58	63	12
	63	68	18
	47	55	22
	34	36	27

5 ЛАБОРАТОРНАЯ РАБОТА № 5. Проведение регрессионного анализа средствами среды программирования R

Цель работы: Изучение основ регрессионного анализа.

Основные теоретические положения

При исследовании взаимосвязей между выборками помимо корреляции различают также и **регрессию**. Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или более независимых переменных. Соответственно, наряду с корреляционным анализом еще одним инструментом изучения стохастических зависимостей является регрессионный анализ. Регрессионный анализ устанавливает формы зависимости между случайной величиной Y (зависимой) и значениями одной или нескольких переменных величин (независимых), причем значения последних считаются точно заданными. Такая зависимость обычно определяется некоторой математической моделью (уравнением регрессии), содержащей несколько неизвестных па-

раметров. В ходе регрессионного анализа на основании выборочных данных находятся оценки этих параметров, определяются статистические ошибки оценок или границы доверительных интервалов и проверяется соответствие (адекватность) принятой математической модели экспериментальным данным.

В **MS Excel** экспериментальные данные аппроксимируются линейным уравнением до 16 порядка.

Для получения коэффициентов регрессии используется процедура **Регрессия** из **Пакета анализа**. Кроме того, могут быть использованы функция **ЛИНЕЙН** для получения параметров регрессионного уравнения и функция **ТЕНДЕНЦИЯ** для получения предсказанных значений Y в требуемых точках.

Для реализации процедуры **Регрессия** необходимо:

- выполнить команду **Сервис – Анализ данных**;
- в появившемся диалоговом окне Анализ данных в списке **Инструменты анализа** выбрать строку **Регрессия**;
- в появившемся диалоговом окне задать **Входной интервал Y**, то есть ввести ссылку на диапазон анализируемых зависимых данных, содержащий один столбец данных;
- указать **Входной интервал X**, то есть ввести ссылку на диапазон независимых данных, содержащий до 16 столбцов анализируемых данных;
- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа (рисунок 8);

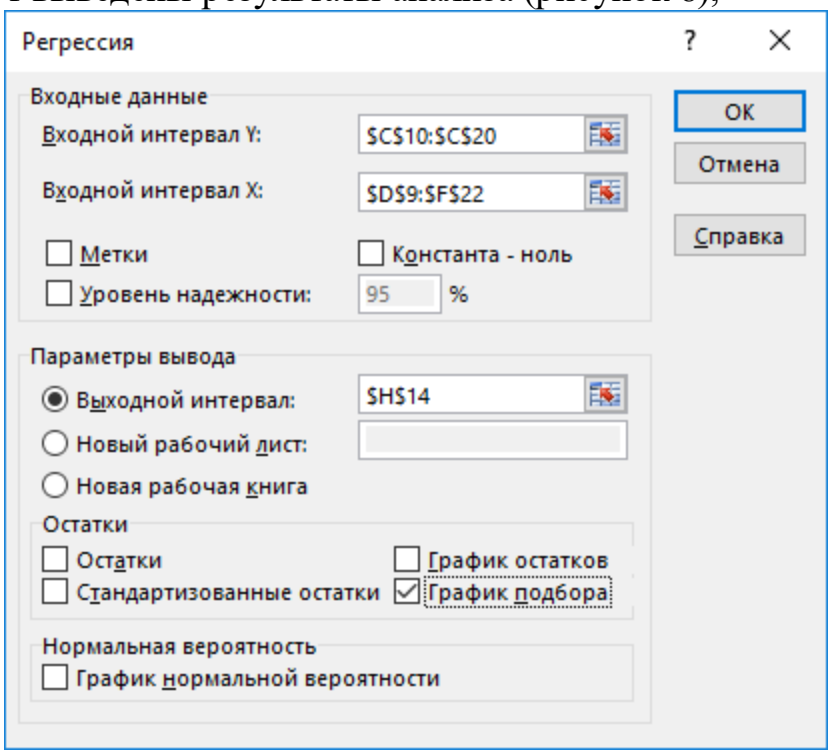


Рисунок 8 – Пример заполнения диалогового окна **Регрессия**

- если необходимо визуально проверить отличие экспериментальных точек от предсказанных по регрессионной модели, следует установить флажок в поле **График подбора**:
- нажать кнопку **ОК**.

Результаты анализа. Выходной диапазон будет включать в себя ре-

зультаты дисперсионного анализа, коэффициенты регрессии, стандартную погрешность вычисления Y, среднеквадратичные отклонения, число наблюдений, стандартные погрешности для коэффициентов.

Интерпретация результата. Значения коэффициентов регрессии находятся в столбце Коэффициенты.

В столбце **P-Значение** приводится достоверность отличия соответствующих коэффициентов от нуля. В случаях, когда $P > 0{,}05$, коэффициент может считаться нулевым; это означает, что соответствующая независимая переменная практически не влияет на зависимую переменную.

Приводимое значение *R*-квадрат (коэффициент детерминации) определяет, с какой степенью точности полученное регрессионное уравнение аппроксимирует исходные данные. Если *R*-квадрат > 0,95, говорят о высокой точности аппроксимации (модель хорошо описывает явление). Если *R*-квадрат лежит в диапазоне от 0,8 до 0,95, говорят об удовлетворительной аппроксимации (модель в целом адекватна описываемому явлению). Если *R*-квадрат < 0,6, принято считать, что точность аппроксимации недостаточна и модель требует улучшения (введения новых независимых переменных, учета нелинейностей и т. д.).

Пример. В отделе снабжения гостиницы имеется информация об изменении стоимости стирального порошка за длительный период времени. Сопоставляя ее с изменениями курса доллара за этот же период времени, можно построить регрессионное уравнение. Ниже (таблица 3) приведены стоимость пачки стирального порошка (в рублях) и соответствующий курс доллара (руб/USD).

Таблица 3 – Информация об изменении стоимости стирального порошка

Номер	1	2	3	4	5	6	7	8
Порошок	5	7	9	12	15	16	20	25
Курс	6,3	9	12	15	19	21	25	29,3

Необходимо на основании этих данных построить регрессионное уравнение, позволяющее по курсу доллара определять предполагаемую стоимость пачки стирального порошка.

Решение.

1. Введите данные в рабочую таблицу: стоимость пачки порошка – в диапазон A1:A8; курс доллара – в диапазон B1:B8.
2. Выполните команду **Сервис – Анализ данных** и выберите строку Регрессия.
3. В появившемся диалоговом окне задайте **Входной интервал *Y*** – это диапазон ячеек A1:A8 (обратите внимание, что зависимые данные – это те данные, которые предполагается вычислять).
4. Также укажите **Входной интервал *X***, задав диапазон независимых данных B1:B8 (независимые данные – это те данные, которые будут измеряться или наблюдаться).
5. Установите флажок в поле **График подбора.**
6. Далее укажите **Выходной диапазон**, например, ячейку C1.
7. Нажмите кнопку **ОК**.

Результаты анализа. В выходном диапазоне появятся результаты и график подбора (рисунок 8).

Интерпретация результатов. В таблице **Дисперсионный ана**лиз оценивается общее качество полученной модели: ее достоверность по уровню значимости критерия Фишера – *p*, который должен быть меньше, чем 0,05 (строка **Регрессия,** столбец **Значимость *F***, в примере – 1,58E-07 (0,000000158), то есть *p* = 0,000000158 и модель значима) и степень точности описания моделью процесса **R-квадрат** (вторая строка сверху в таблице **Регрессионная статистика**, в примере R-квадрат = 0,992). Поскольку R-квадрат > 0,95, можно говорить о высокой точности аппроксимации (модель хорошо описывает явление (рисунок 9).

ВЫВОД ИТОГОВ

ессионная статистика	
Множеств	0,996016
R-квадрат	0,992049
Нормиро	0,990723
Стандарт	0,651083
Наблюде	8

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	1	317,3315	317,3315	748,5832	1,58E-07
Остаток	6	2,543457	0,42391		
Итого	7	319,875			

Коэффициен	*артная о*	*татистич*	*-Значение*	*ижние 95*	*ерхние 95*	*ижние 95,0*	*рхние 95,0%*
Y-пересеч	-0,83091	0,576322	-1,44174	0,199462	-2,24112	0,5793	-2,24112	0,5793
Перемен	0,846613	0,030943	27,36025	1,58E-07	0,770897	0,922328	0,770897	0,922328

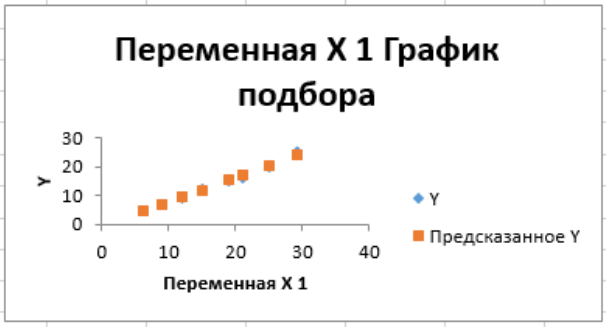


Рисунок 9 – Результаты анализа и график соответствия экспериментальных точек и предсказанных по регрессионной модели

Далее необходимо определить значения коэффициентов модели. Они определяются из таблицы в столбце **Коэффициенты** – в строке **Y-пересечение** приводится свободный член; в строках соответствующих переменных приводятся значения коэффициентов при этих переменных. В столбце **p-значение** приводится достоверность отличия соответствующих коэффициентов от нуля. В случаях, когда *p* > 0,05, коэффициент может считаться нулевым. Это означает, что соответствующая независимая переменная практически не влияет на зависимую переменную и коэффициент может быть убран из уравнения.

Отсюда выражение для определения стоимости пачки порошка в рублях будет иметь следующий вид: **-0,83 + 0,847*(Курс доллара. руб./USD).**

Полученная модель с высокой точностью позволяет определять стоимость пачки стирального порошка (R^2 = 99,2 %).

Воспользовавшись полученным уравнением, можно рассчитать ожидаемую стоимость пачки стирального порошка при изменениях курса доллара. Например, при курсе доллара 35 руб./USD ожидаемая стоимость пачки порошка равна 28,8 руб.

Рассмотрим более подробно средства **MS Excel** для построения уравнения регрессии. Пусть Вы являетесь менеджером фирмы по продажам подержанных автомобилей и постоянно ведете учет проданных автомобилей. В вашем распоряжении имеются две наблюдаемые величины: x – номер недели, y – число проданных за неделю автомобилей (таблица 4). Фирма совсем молодая, была создана шесть недель назад, и поэтому в вашем распоряжении имеется статистика только за этот весьма ограниченный промежуток времени.

Таблица 4 – Значения наблюдаемых величин

Наблюдаемые величины	**Значения**					
x	1	2	3	4	5	6
y	7	9	12	13	14	17

Вы хотите сначала смоделировать ту динамику продаж, которая имеет место, а на основе построенной модели затем попытаться заглянуть в будущее, т. е. спрогнозировать ожидаемый объем продаж на ближайшие недели.

В качестве модели вы решили взять простейшую модель $y = mx + b$, наилучшим образом описывающую наблюдаемые значения. Обычно m и b подбираются так, чтобы минимизировать сумму квадратов разностей теоретических и наблюдаемых значений зависимой переменной (y), т. е. минимизировать:

$$z = \sum_{i=1}^{n}(y_i - mx_i - b)^2 ,$$

где n – число наблюдений (в данном случае n = 6).

Для решения этой задачи необходимо выполнить следующие действия:

1. Заполнить ячейки A2:B7 (рисунок10).
2. Отвести под переменные m и b ячейки D2 и E2.
3. В ячейку F2 ввести минимизируемую функцию (это формула массива, поэтому не забудьте завершить ее ввод нажатием комбинации клавиш <Shift> + <Ctrl> + <Enter>):

   {=СУММ((B2:B7 - D2 * A2:A7 - E2) ^ 2)}
4. Выполнить команду **Сервис – Поиск решения**. Диалоговое окно **Поиск решения** следует наполнить, как покатано на рисунке 10. Отметим, что на переменные m и b не налагается никаких ограничений.
5. Нажать кнопку **Выполнить**. В результате вычислений средство **Поиск решения** найдет m = 1,88571 и b = 5,40 (рисунок 10).

C7 | fx | =$D$5*A7+$E$5

	A	B	C	D	E	F
1	X	Y	Теоретические значения y	m	b	Функция цели
2	1	7	7	1,885714	5,4	1,771428571
3	2	9	9			
4	3	12	11	Наклон	Отрезок	
5	4	13	13	1,885714	5,4	
6	5	14	15			
7	6	17	17			
8						

Рисунок 10 – Теоретическое значение наблюдаемой величины и коэффициенты уравнения регрессии

Функции MS Excel для построения линейного уравнения регрессии. Рассмотрим некоторые функции категории **Статистические** для построения уравнения регрессии. Параметры *m* и *b* линейной модели **y = *mx* + *b*** из предыдущего примера можно определить при помощи функций **НАКЛОН** и **ОТРЕЗОК**.

Функция **НАКЛОН** определяет коэффициент наклона линейного тренда, а функция **ОТРЕЗОК** – точку пересечения линии линейного тренда с осью ординат.

Линейная зависимость между переменными описывается уравнением общего вида $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n + \varepsilon$ где y - зависимая переменная, $\alpha_0, \alpha_1, ..., \alpha_n$ - неизвестные константы, $x_1, x_2, ..., x_n$ - известные (независимые) переменные, и ε - нормально распределенная случайная величина с нулевым матожиданием и дисперсией σ_{err}^2. Задачей построения линейной среднеквадратической модели регрессионной зависимости переменной y от независимых переменных является получение оценки параметров $\alpha_0, \alpha_1, ..., \alpha_n$ и оценка адекватности построенной модели вида

$$\hat{y} = a_0 + a_1 x_1 + ... + a_n x_n$$

где $a_0, a_1, ..., a_n$ - оценки параметров $\alpha_0, \alpha_1, ..., \alpha_n$.

Рассмотрим простейший случай одной независимой переменной:

$\hat{y} = a + bx$ В этом уравнении модели линейной регрессии a - свободный член, а параметр b определяет наклон линии регрессии по отношению к осями координат. Параметры a и b определяются методом наименьших квадратов, который приводит к формуле:

$$b = r_{xy} \frac{S_y}{S_x}, \; a = \bar{y} - b\bar{x},$$

где

$\bar{y}, \bar{x}$ - выборочные средние арифметические;

S_x, S_y - выборочные средние квадратичные отклонения;

r_{xy} - выборочный коэффициент корреляции.

Для построения линейной модели регрессии используется функция **lm(formula=f)**, которая в простейшем случае содержит только формулу от переменных (векторов, содержащих элементы парной выборки); запись *y~x* означает, что строится модель зависимости *y* от *x*.

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.8,6.4)
> p.lm<-lm(formula=x~y)
> summary(p.lm)

Residuals:
      1       2       3       4       5
-1.7151 -0.3409  2.5529 -2.3954  1.8985

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0845     3.5050   1.165    0.328
y             0.4558     0.4985   0.914    0.428

Residual standard error: 2.511 on 3 degrees of freedom
Multiple R-Squared: 0.2179,     Adjusted R-squared: -
0.0428
F-statistic: 0.8358 on 1 and 3 DF,  p-value: 0.428
```

Команда **summary()** выдает полную информацию о построенной модели:

значения остатков (residuals - разность модельных и истинных значений переменной *y*). Если объем выборки большой, то печатается оценка распределения остатков (квартили).

коэфициенты модели и оценку их значимости по критерию Стьюдента (в нашем случае все коэфициенты не значимы, поскольку все вероятности (0.328 и 0.428) больше 0.05 - т.е. нельзя считать, что существует линейная зависимость между x и *y*).

Оценку значимости зависимости по критерию Фишера и квадрат коэфициента корреляции (R-squared), который показывает долю дисперсии *y*, объясненной с использованием модели (исправленное значение для R2 равно 0, статистика Фишера *F*=0.8358, уровень значимости критерия Фишера 42.8%, т.е. зависимость отсуствует).

Для визуализации построенной модели можно использовать вспомогательные функции:

Описание функций

```
abline(a, b, untf = FALSE, ...)
abline(h=, untf = FALSE, ...)
abline(v=, untf = FALSE, ...)
```

Параметры

`a,b`	Параметры в линейном уравнении
`untf`	Если TRUE, то рисует линию в преобразованных координатах
`h,v`	Y и X значения для горизонтальной и вертикальной линии соответственно

```
plot(x, y, xlim=range(x),ylim=range(y),type="p",
main, xlab, ylab, ...)
```

Параметры

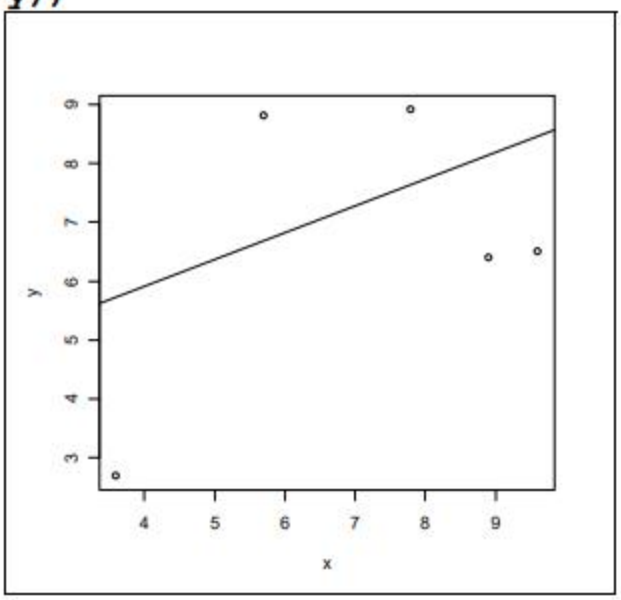
`X,Y`	Координаты точек x и y.
`xlim, ylim`	Значения для осей x и y.
`Type`	Тип графика(“ p” для точек)
`Main`	**Название графика**
`Xlab,ylab`	Название осей.

Функция **`abline()`** строит прямую по найденным a и b.

Функция **`plot()`** строит экспериментальные точки.

Пример

```
plot(x,y)
abline(lm(x~y))
```



Задания для лабораторной работы

1 Построить зависимость зарплаты (р.) от возраста сотрудника гостиницы по следующим данным:

Вариант №	Данные						
	Возраст	20	50	45	40	25	30

1	Зарплата	800	2500	2500	2000	1200	1800
2	Зарплата	900	2600	2500	2200	1300	1900
3	Зарплата	700	2700	2500	2300	1100	1700
4	Зарплата	1000	2400	2300	1900	1200	1500
5	Зарплата	900	2400	2400	2000	1400	1700
6	Зарплата	1000	2500	2300	2100	1200	1900
7	Зарплата	700	2500	2400	1900	1300	1600
8	Зарплата	900	2300	2300	1900	1200	1800
9	Зарплата	800	2600	2600	2100	1300	1700
10	Зарплата	800	2600	2500	2200	1200	1800

2 Построить зависимость жизненной емкости легких в литрах (Y) от роста в метрах (X_1) и возраста в годах (X_2) для группы мужчин:

Вариант №	Данные		
	X_1	X_2	Y
1	1,85	18	5,4
	1,8	25	6,7
	1,75	20	4,8
	1,7	24	5,1
	1,68	21	4,5
	1,73	19	4,8
	1,77	22	5,11
	1,81	23	5,6
	1,76	18	4,7
2	1,85	20	6,2
	1,8	18	4,7
	1,75	18	5,4
	1,7	21	5,6
	1,68	25	4,7
	1,73	21	4,7
	1,77	24	5,8
	1,81	25	5
	1,76	17	5,8
3	1,85	22	5,1
	1,8	18	5,7
	1,75	23	5,3
	1,7	20	6
	1,68	20	5,4
	1,73	23	6,3
	1,77	19	5,7
	1,81	22	5,7
	1,76	21	4,7
4	1,85	19	5,2
	1,8	18	5,7
	1,75	18	4,7
	1,7	21	5,3
	1,68	21	6,3
	1,73	23	4,7
	1,77	24	5,1
	1,81	17	4,9

	1,76	20	5,3
5	1,85	21	5,4
	1,8	21	6,1
	1,75	24	5
	1,7	25	6
	1,68	18	5,5
	1,73	23	5
	1,77	22	5,4
	1,81	23	6,2
	1,76	18	6
6	1,85	19	6,2
	1,8	18	6
	1,75	22	6,1
	1,7	19	5,6
	1,68	25	5,7
	1,73	22	5,5
	1,77	17	5,4
	1,81	21	4,9
	1,76	20	4,6
7	1,85	17	4,8
	1,8	23	4,5
	1,75	23	5,6
	1,7	21	5,6
	1,68	18	4,9
	1,73	17	5,1
	1,77	24	5,4
	1,81	19	5,6
	1,76	17	4,5
8	1,85	18	6,2
	1,8	24	5,5
	1,75	23	4,7
	1,7	20	5,2
	1,68	22	4,9
	1,73	25	6,1
	1,77	18	5,8
	1,81	19	4,8
	1,76	20	5,2
9	1,85	19	5,2
	1,8	23	5,5
	1,75	25	4,7
	1,7	25	4,8
	1,68	23	5,1
	1,73	23	5,5
	1,77	21	4,7
	1,81	25	5,8
	1,76	21	6
10	1,85	25	4,9
	1,8	25	6
	1,75	19	4,6
	1,7	22	5,4

	1,68	21	6,2
	1,73	25	5,9
	1,77	20	4,6
	1,81	24	4,8
	1,76	20	4,8

3 Определить должное значение жизненной емкости легких для мужчины возраста 22-х лет и роста 183 см из регрессионного уравнения, полученного в предыдущем упражнении.

4 Имеются данные о цене на нефть x (ден. ед.) и индексе акций нефтяных компаний y (усл. ед.):

Вариант №	Данные						
1	x	17,28	17,05	18,30	18,80	19,20	18,50
	y	537	534	550	555	560	552
2	x	18,71	19,44	18,65	17,18	17,77	17,39
	y	556,3	555,94	544,9	540,72	541,07	559,47
3	x	18,99	17,71	18,82	17,42	18,4	18,15
	y	549,66	555,85	541,18	544,68	537,08	538,97
4	x	19,11	18,72	18,56	18,99	18,77	18,77
	y	559,48	552,02	533,85	539,22	549,42	553,12
5	x	17,86	19,29	17,58	18,83	17,75	17,49
	y	537,66	535,68	541,28	546,18	550,84	551,58
6	x	18,42	18,15	18,97	17,61	17,91	18,96
	y	557,44	535,81	552,21	540,84	536,83	548,6
7	x	18,33	17,2	17,39	18,97	17,94	18,59
	y	555,61	544,63	549,14	558,76	554,71	549,29
8	x	17,57	19,31	18,4	17,44	17,58	17,12
	y	557,69	539,13	559,75	557,16	536,89	537,44
9	x	19,3	18,84	17,92	17,87	17,98	18,13
	y	530,91	559,49	550,67	548,35	536,02	536,65
10	x	17,11	18,17	19,04	18,85	18,71	17,52
	y	532,44	548,24	549,15	534,32	530,48	543,26

Построить зависимость индекса акций нефтяных компаний от цены на нефть.

6 ЛАБОРАТОРНАЯ РАБОТА № 6. Анализ временных рядов и прогнозирование средствами среды программирования R

Цель работы: Изучение моделей временных рядов и методов их анализа.

Основные теоретические положения

Временной ряд – это последовательность наблюдений, упорядоченная по времени: $y_1, y_2, ..., y_t, ..., y_n$, где y_t – числа, представляющие наблюдения некоторой переменной в n равноотстоящих моментах времени $t = 1, 2, ..., n$.

Примеры данных, которые необходимо изучать во времени: цены на товар, деловая активность, национальный валовой продукт.

Особенность временных рядов – зависимость данных, характер которой

может определяться положением наблюдений в последовательности.

Основные задачи анализа временных рядов:

- прогнозирование на основе знания прошлого;
- сжатое описание характерных особенностей ряда;
- управление процессом, порождающим ряд.

В анализе временных рядов предполагается, что исходные данные содержат детерминированную и случайную (ε_t) составляющие. В общем случае детерминированная составляющая может быть представлена в виде совокупности следующих компонентов:

- *тренда* u_t, определяющего главную тенденцию временного ряда;
- *циклов* (*циклической составляющей*) W_t – более или менее регулярных колебаний относительно тренда;
- *сезонной составляющей* S_t – периодических колебаний.

Временной ряд может быть представлен различными математическими моделями.

Аддитивная модель

$$y_t = u_t + W_t + S_t + \varepsilon_t .$$

Мультипликативная модель

$$y_t = u_t W_t S_t \varepsilon_t .$$

Если предположить, что сезонная составляющая S_t пропорциональна сумме тренда и циклической составляющей $S_t = (u_t + W_t)C_t$, то временной ряд будет представлен в виде *смешанной модели*:

$$y_t = (u_t + W_t)(1 + C_t) + \varepsilon_t .$$

Выбор модели зависит от конкретной совокупности явлений, определяющих данный временной ряд, и их взаимосвязей.

Основные цели.

Существует две основные цели анализа временных рядов:

1) определение природы ряда.

2) прогнозирование (предсказание будущих значений временного ряда по настоящим и прошлым значениям).

Обе эти цели требуют, чтобы модель ряда была идентифицирована и более или менее формально описана. Как только модель определена, можно с ее помощью интерпретировать рассматриваемые данные. Не обращая внимания на глубину понимания и справедливость теории, можно затем экстраполировать ряд на основе найденной модели, т. е. предсказать его будущие значения.

Для работы подключим следующие пакеты:

```
library("lubridate") # работа с датами
library("zoo") # работа с временными рядами
library("xts") # дополнительные функции для работы с временными рядами
library("dplyr") # работа с наборами данных
library("ggplot2") # графики
library("forecast") # прогнозы
library("lmtest")# тестирование гипотез в линейных моделях
library("quantmod") # загрузка данных с различных источников
```

Для начала посмотрим, как сгенерировать тестовые выборки. Для этого воспользуемся функцией `arima.sim` из базового пакета `stats`.

Выполним симуляцию AR(1)-процесса в 100 наблюдений по модели: $y_t = 0.6 y_{t-1} + \varepsilon_t$ и поместим в переменную `y`.

```
y <- arima.sim(n=100, list(ar=0.6))
```

В этой функции необходимо указать объем выборки, а также в параметре `list` коэффициенты модели.

Можно построить полученный ряд на графике (рис. 17):

```
plot(y)
```

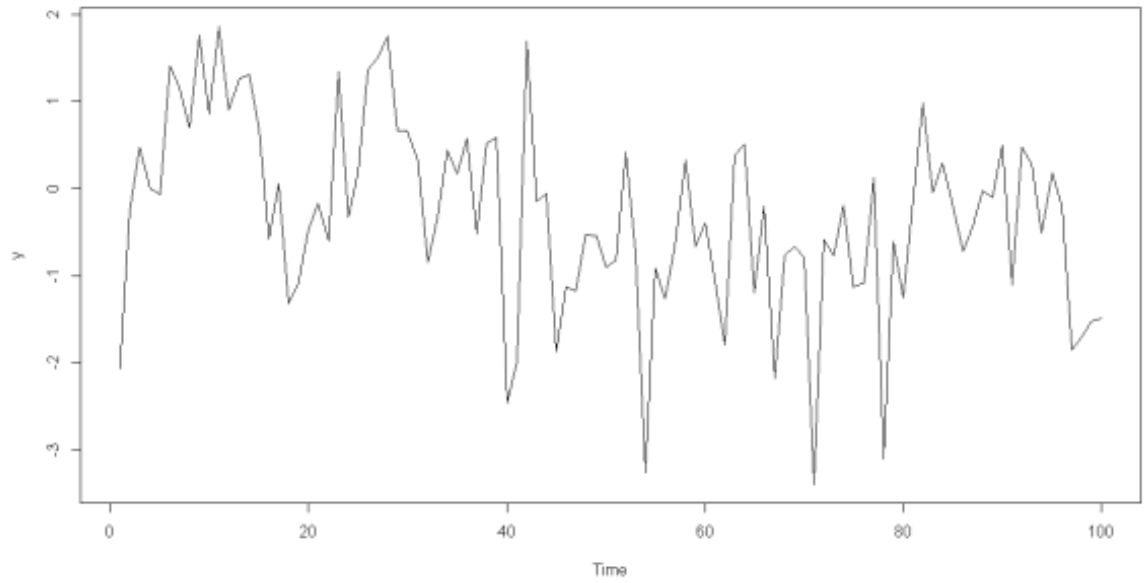


Рис. 17. График временного ряда

Графики автокорреляционной и частной автокорреляционной функций для процесса `y` вызываются командами соответственно:

```
Acf(y)
Pacf(y)
```

Все три графика одновременно могут быть вызваны командой:

```
tsdisplay(y)
```

Результат выполнения команды показан на рисунке 18.

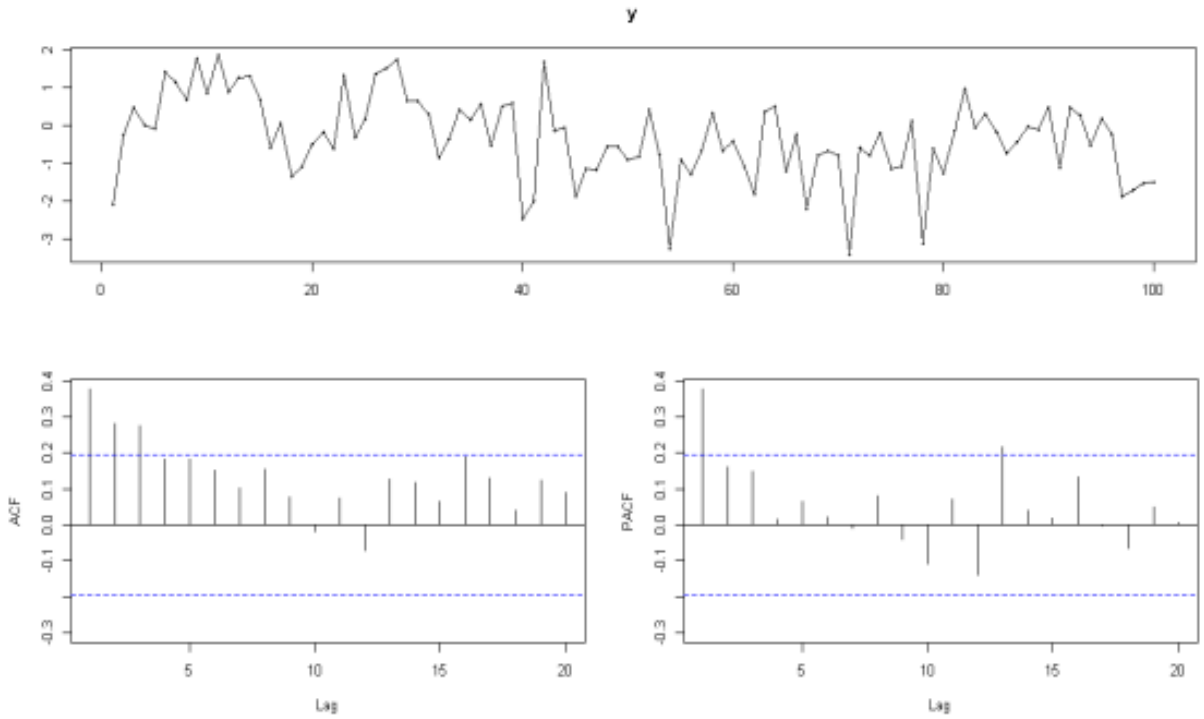


Рис. 18. График временного ряда с графиками автокорреляционной и частной автокорреляционной функций

Посмотрим еще несколько примеров симуляции AR, MA, ARMA, ARIMA процессов.

Процесс MA(1) $y_t = \varepsilon_t - 0.8\varepsilon_{t-1}$

```
arima.sim(n=100, list(ma=-0.8))
```

Процесс ARMA(1,1) $y_t = 0.5y_{t-1} + \varepsilon_t - 0.8\varepsilon_{t-1}$

```
arima.sim(n=100, list(ma=-0.8, ar=0.5))
```

Процесс ARMA(2,2) $y_t = 0.9y_{t-1} - 0.5y_{t-2} + \varepsilon_t - 0.2\varepsilon_{t-1} + 0.3\varepsilon_{t-2}$

```
arima.sim(n = 100, list(ar = c(0.9, -0.5), ma = c(-0.2, 0.3))
```

Процесс случайного блуждания $y_t = y_{t-1} + \varepsilon_t$

```
arima.sim(n=100, list(order=c(0,1,0)))
```

Белый шум $y_t = \varepsilon_t$:

```
arima.sim(n=100, list(order=c(0,0,0)))
```

Попробуем подобрать различные ARIMA-модели под реальные данные. Для примера выберем временной ряд, содержащий 98 еже-

годных наблюдений за уровнями воды в озере Гурон с 1875 по 1972 гг. (рис. 19).

```
help(LakeHuron)
```

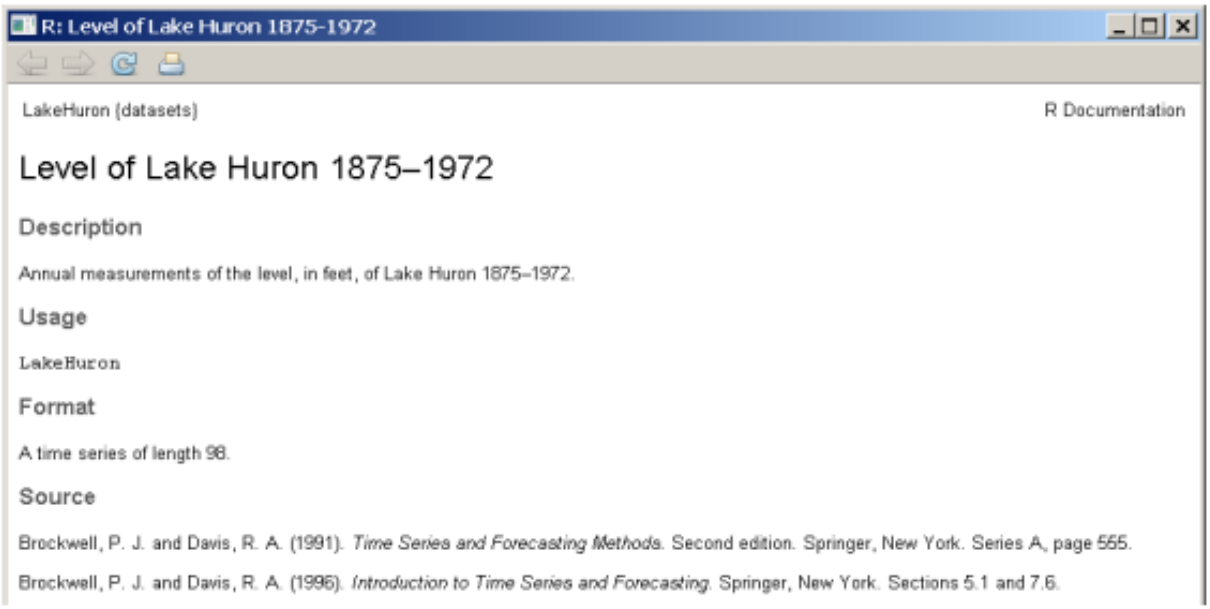


Рис. 19. Справка по набору данных `LakeHuron`

Оценим ARMA(2,1) модель с помощью функции `Arima` из пакета `forecast`:

```
mod<- Arima(y, order=c(2,0,1))
```

Посмотрим на результат оценивания:

```
summary(mod)
Series: y
ARIMA(2,0,1) with non-zero mean
Coefficients:
         ar1      ar2     ma1  intercept
0.7829  -0.0342  0.2857   579.0533
s.e.  0.3262   0.2845  0.3144     0.3467
sigma^2 estimated as 0.4749:  log likelihood=-103.24
AIC=216.48   AICc=217.13   BIC=229.4
Training set error measures:
                     ME      RMSE       MAE          MPE       MAPE      MASE
Training set -0.008638553 0.6891058 0.5492008 -0.001635858 0.09486097 0.9378958
                    ACF1
Training set 0.002260998
```

Получили модель $y_t = 579.0533 + 0.7829 y_{t-1} - 0.0342 y_{t-2} + \varepsilon_t + 0.2857 \varepsilon_{t-1}$. В строке `s.e.` перечислены стандартные ошибки коэффициентов. Оценка дисперсии $\hat{\sigma}^2 = 0.4749$. Строкой ниже перечислены значения критерия Акаике (AIC=216.48), скорректированного критерия Акаике (AICc=217.13) и критерия Шварца (BIC=229.4).

Можно отдельно вывести значение критерия Акаике командой:

```
AIC(mod)
```

Командой `fitted(mod)` можно получить модельные значения временного ряда. Построим реальные и модельные значения на одном графике (рис. 20):

```
matplot(cbind(y,fitted(mod)),type='l')
```

Функция `matplot` строит значения столбцов матрицы, поэтому ее аргументом нужно задать столбцы. Команда `cbind` соединяет в матрицу столбец *y* и столбец значений, рассчитанных по модели. `type='l'` указывает, что тип графика — линия.

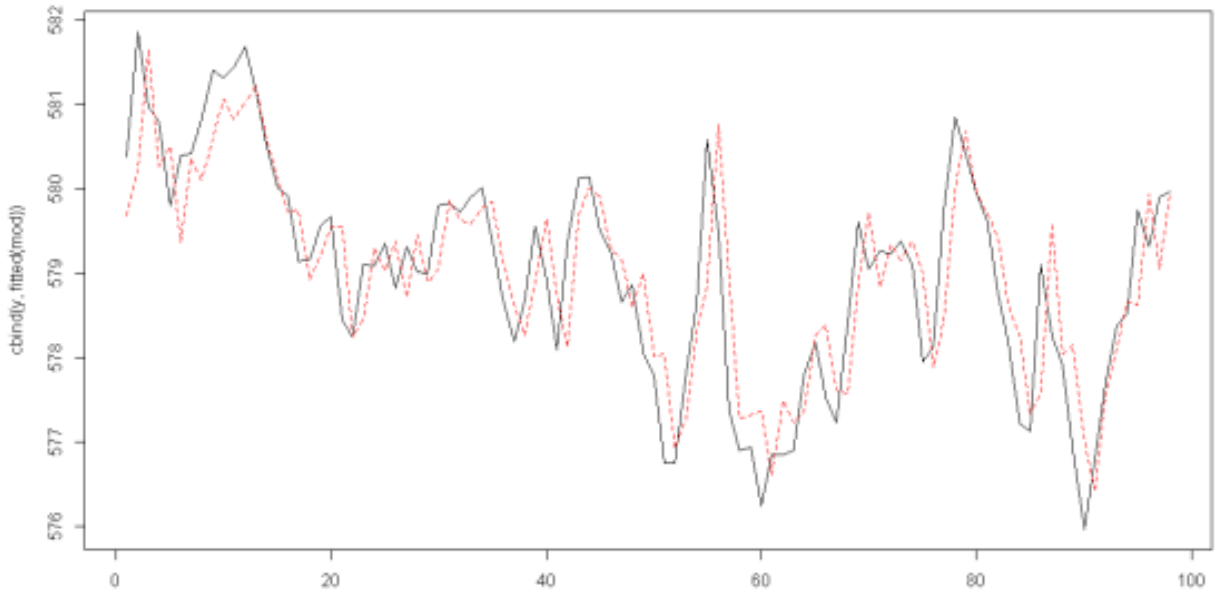


Рис. 20. График модельных и реальных значений временного ряда

Построим прогноз на 5 шагов вперед:

```
prognoz<- forecast(mod, h=5)
```

Выведем полученные значения. Результат включает 80% и 95% доверительный интервал для прогноза:

```
prognoz
     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
1973       579.7407 578.8576 580.6238 578.3901 581.0913
1974       579.5605 578.2680 580.8530 577.5838 581.5372
1975       579.4269 577.9528 580.9009 577.1725 581.6813
1976       579.3284 577.7645 580.8924 576.9366 581.7203
1977       579.2559 577.6453 580.8666 576.7927 581.7192
```

Строим график прогноза (рис. 21):

```
plot(prognoz)
```

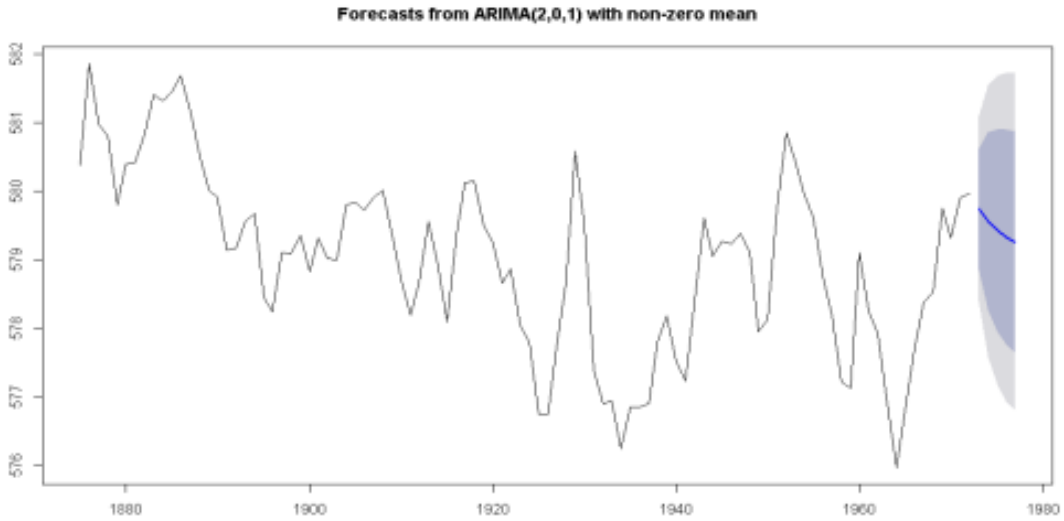


Рис. 21. График прогноза по модели ARIMA

R также может автоматически подбирать ARIMA-модель по штрафному критерию (минимальное значение критерия Акаике):

```
mod_a<- auto.arima(y)
summary(mod_a)
Series: y
ARIMA(0,1,4) with drift
Coefficients:
         ma1      ma2      ma3      ma4    drift
0.0584  -0.3158  -0.3035  -0.2349  -0.0205
s.e.  0.1031   0.1058   0.1100   0.1299   0.0167
sigma^2 estimated as 0.4806:  log likelihood=-101.09
AIC=214.18   AICc=215.12   BIC=229.63
Training set error measures:
                     ME       RMSE        MAE             MPE      MAPE
MASE
Training  set  0.002898915  0.6886758  0.5441565  0.0003830464  0.093983
0.9292814
                   ACF1
Training set 0.01081339
```

«ARIMA(0,1,4) with drift» означает, что была подобрана ARIMA(0,1,4) с линейным трендом (drift), но без константы (отсутствует параметр intercept).

Таким образом, получим модель

$$y_t = y_{t-1} + \varepsilon_t + 0.0584\varepsilon_{t-1} - 0.3158\varepsilon_{t-2} - 0.3035\varepsilon_{t-3} - 0.2349\varepsilon_{t-4} - 0.0205t.$$

Загрузка данных из различных источников в R

Для корректного перевода дат на английский язык в русскоязычной Windows, необходимо сначала выполнить следующую команду:

```
Sys.setlocale("LC_TIME","C")
```

Пакет `quantmod` позволяет загружать данные из нескольких источников, а именно:

– Yahoo! Finance (OHLC data) — http://finance.yahoo.com/;

– Federal Reserve Bank of St. Louis FRED® (11,000 экономических временных рядов) — http://research.stlouisfed.org/fred2/;

– Google Finance (OHLC data) — http://finance.google.com/;

– Oanda, The Currency Site (FX and Metals) — http://www.oanda.com/.

Для загрузки данных используется одна и та же функция `getSymbols`, например:

```
#данные о стоимости акций Гугл с finance.google.com за период с 1 января 2015г. по 1 декабря 2015г.
getSymbols("GOOG", src="google", from="2015-01-01", to="2015-12-01")
```

GOOG — это краткое название акций компании на бирже (ти́кер). Теперь эти данные загружены в переменную с таким же названием `GOOG`.

Если не указывать начальную или конечную дату, то будут загружены все доступные данные. Можно посмотреть шесть первых значений и шесть последних значений командами `head` и `tail` соответственно.

```
head(GOOG)
```

	GOOG.Open	GOOG.High	GOOG.Low	GOOG.Close	GOOG.Volume	GOOG.Adjusted
2015-01-02	529.0124	531.2724	524.1023	524.8124	1447600	524.8124
2015-01-05	523.2624	524.3324	513.0623	513.8723	2059800	513.8723
2015-01-06	515.0024	516.1773	501.0523	501.9623	2899900	501.9623
2015-01-07	507.0023	507.2463	499.6522	501.1023	2065100	501.1023
2015-01-08	497.9922	503.4823	491.0022	502.6823	3353600	502.6823
2015-01-09	504.7623	504.9223	494.7922	496.1723	2069400	496.1723

```
tail(GOOG)
```

	GOOG.Open	GOOG.High	GOOG.Low	GOOG.Close	GOOG.Volume	GOOG.Adjusted
2015-11-23	757.45	762.708	751.82	755.98	1414500	755.98
2015-11-24	752.00	755.279	737.63	748.28	2333100	748.28
2015-11-25	748.14	752.000	746.06	748.15	1122100	748.15
2015-11-27	748.46	753.410	747.49	750.26	838500	750.26
2015-11-30	748.81	754.930	741.27	742.60	2035300	742.60
2015-12-01	747.11	768.950	746.70	767.04	2129900	767.04

Посмотрим на график (рис. 22):

```
barChart(GOOG, theme = "white")
```

Рис. 22. График стоимости акций Гугл GOOG

Также можно посмотреть на цену закрытия с помощью функции `tsdisplay` (рис. 23):

```
tsdisplay(GOOG$GOOG.Close)
```

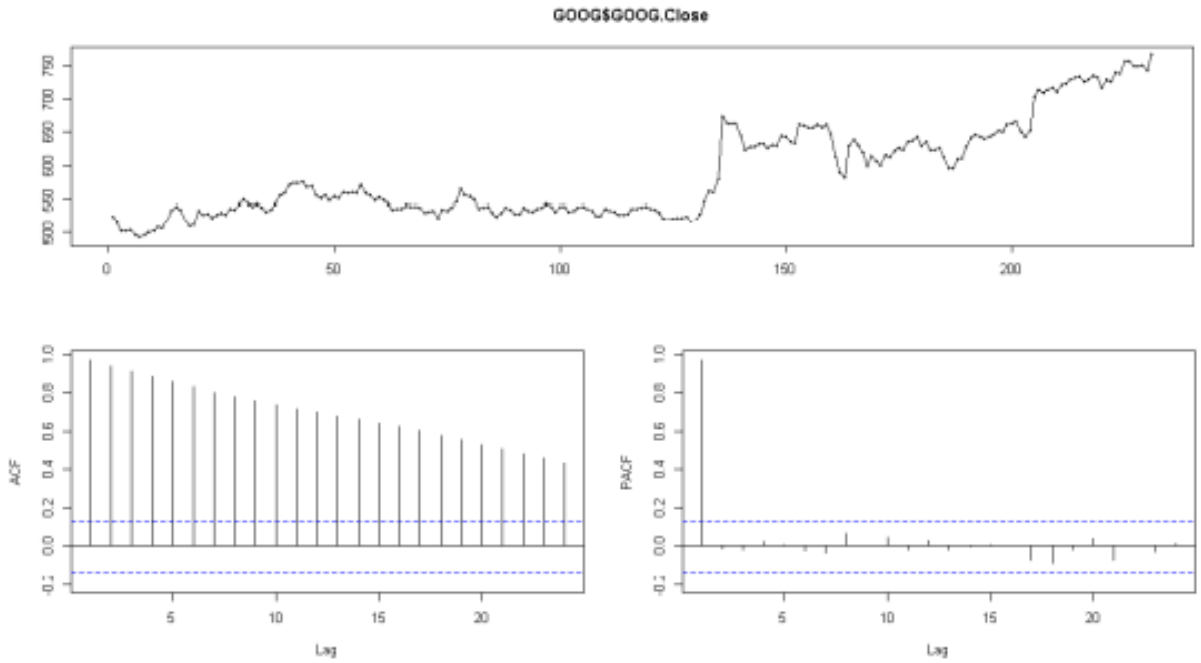


Рис. 23. График цены закрытия акций Гугл GOOG

Приведем еще несколько примеров:

```
#данные о стоимости акций Yahoo с finance.google.com
getSymbols("YHOO",src="google")
#данные о стоимости акций Apple с finance.yahoo.com
getSymbols("AAPL",src="yahoo", from="2015-01-01", to="2015-12-01")
```

Японские свечи (рис. 24):

```
candleChart(AAPL, multi.col=TRUE, theme="white")
```

Рис. 24. График стоимости акций Apple AAPL

Для более детальной информации можно обратиться на сайт разработчиков пакета [7].

Практическое задание

Положение на рынке местного производителя продукта П1 представляют следующие значения (исходные данные по кварталам определить самостоятельно) (таблица 6.1).

Таблица 6.1 – Шаблон таблицы для исходных данных

Продажи за квартал				
Год	I	II	III	IV
1				
2				
3				
4				
5				

Пример заполнения шаблона таблицы представлен в таблице 3.2.

Таблица 3.2 – Пример исходных данных для решения практического задания

Продажи за квартал				
Год	I	II	III	IV
1	19	24	38	25
2	21	28	44	23
3	23	31	41	23
4	24	35	48	21
5	22	37	50	22

1 Вычислите сезонные индексы для данных из таблицы 3.2 (используйте центрированные средние значения за четыре квартала).

2 Исключите сезонную составляющую из этих данных.

3 Методом наименьших квадратов найдите параметры прямой, которая наилучшим образом характеризует основную тенденцию временного ряда в данных о продаже продукта.

4 Определите циклический компонент в этом временном ряду, исключив тренд из исходных данных.

Контрольные вопросы

1. В чем отличие автокорреляционной и частной автокорреляционной функций?
2. Что такое случайное блуждание и белый шум?
3. Как можно сравнить модели временного ряда по точности?

Список литературы

1 Борздова Т. В. Основы статистического анализа и обработка данных с применением Microsoft Excel: учеб. пособие / Т. В. Борздова. – Минск : ГИУСТ БГУ, 2011. – 75 с.

2 Маталыцкий, М.А. Теория вероятностей, математическая статистика и случайные процессы [Электронный ресурс] : учеб. пособие / М.А. Маталыцкий, Г.А. Хацкевич. – Минск: Выш. шк., 2012. – 720 с.: ил.

3 Статистические методы анализа данных: Учебник / Л.И. Ниворожкина, С.В. Арженовский, А.А. Рудяга [и др.]; под общ. ред. д-ра экон. наук, проф. Л.И. Ниворожкиной. — М.: РИОР: ИНФРА-М, 2016. — 333 с.

4 Основы статистического анализа. Практ. по стат. мет. и исслед. операций с исп. пакетов STATISTICA и EXCEL: Уч.пос./ Э.А.Вуколов - 2 изд., испр. и доп. - М.: Форум:НИЦ Инфра-М, 2013. - 464 с.

5 Лекции по случайным процессам: учебное пособие / А. В. Гасников [и др.] ; под ред. А. В. Гасникова. – Москва : МФТИ, 2019. – 285 с.

6 Маталыцкий, М. А. Теория вероятностей, математическая статистика и

случайные процессы: учебное пособие / М. А. Маталыцкий, Г. А. Хацкевич. – Минск: Вышэйшая школа, 2012. – 720 с.

7 Кацман, Ю. Я. Теория вероятностей, математическая статистика и случайные процессы: учебник / Ю. Я. Кацман. – Томск: Томск. политех. ун-т, 2013. – 131 с.

8 Сердобольская, М. Л. Теория случайных процессов. Конспект лекций [Электронный ресурс]: электронное учебное пособие / М. Л. Сердобольская. – Москва, 2011. – Режим доступа: http://cmp.phys.msu.ru/ru/study/math/tsp. – Дата доступа: 15.01.2022.

9 Храмов, А. Г. Теория случайных процессов. Конспект лекций: учебное пособие / А. Г. Храмов. – Самара, 2011.

10 Астахов, В. Н. Теория случайных процессов: учебное пособие / В. Н. Астахов, Г. С. Буланов. – Краматорск: ДГМА, 2006. – 52 с.

Задания для самостоятельной работы

Выполните задания 1-26, теоретические сведения к которым находятся в папке “Теоретические сведения”.

1. Построить эмпирические функции распределения (относительные и накопленные частоты) для роста (в см) 20 человек:

Вариант №	Данные
1	181, 169, 178, 178, 171, 179, 172, 181, 179, 168, 174, 167, 169, 171, 179, 181, 181, 183, 172, 176.
2	193, 190, 187, 179, 182, 184, 182, 183, 190, 188, 186, 184, 184, 186, 178, 191, 185, 186, 183, 188.
3	174, 177, 168, 165, 178, 170, 169, 174, 173, 178, 170, 171, 176, 176, 172, 169, 173, 176, 178, 168.
4	166, 159, 163, 160, 168, 164, 162, 169, 165, 163, 165, 167, 170, 166, 160, 168, 169, 162, 163, 168.
5	186, 182, 180, 192, 179, 182, 183, 186, 184, 189, 187, 186, 190, 182, 185, 184, 187, 186, 191, 183.
6	174, 172, 170, 177, 181, 178, 169, 174, 173, 175, 176, 172, 174, 177, 179, 176, 174, 172, 173, 179.
7	175, 177, 178, 183, 173, 180, 175, 182, 177, 181, 180, 179, 176, 178, 180, 180, 174, 175, 187, 185.
8	170, 189, 176, 181, 182, 179, 183, 172, 188, 171, 190, 185, 181, 177, 176, 182, 187, 177, 175, 178.
9	185, 175, 173, 170, 171, 172, 183, 186, 184, 177, 175, 170, 176, 174, 182, 184, 179, 177, 171, 180.
10	171, 180, 195, 175, 169, 173, 182, 163, 180, 178, 177, 184, 180, 171, 177, 182, 180, 174, 181, 179.

2. Найти распределение по абсолютным частотам для следующих результатов тестирования в баллах:

Вариант №	Данные
1	79, 85, 78, 85, 83, 81, 95, 88, 97 (используйте границы интервалов 70, 79, 89).
2	84, 77, 80, 91, 95, 76, 63, 80, 95 (используйте границы интервалов 70, 79, 89).
3	92, 82, 74, 86, 77, 83, 72, 74, 84 (используйте границы интервалов 70, 79, 89).
4	86, 77, 95, 92, 74, 80, 73, 87, 79 (используйте границы интервалов 70, 79, 89).
5	72, 88, 82, 94, 74, 85, 90, 76, 85 (используйте границы интервалов 70, 79, 89).
6	96, 73, 83, 87, 80, 67, 82, 93, 88 (используйте границы интервалов 70, 79, 89).
7	90, 75, 82, 85, 81, 90, 69, 67, 67 (используйте границы интервалов 70, 79, 89).
8	86, 85, 77, 74, 95, 75, 89, 77, 90 (используйте границы интервалов 70, 79, 89).
9	70, 77, 70, 93, 71, 88, 81, 73, 79 (используйте границы интервалов 70, 79, 89).
10	94, 88, 94, 81, 97, 70, 89, 89, 75 (используйте границы интервалов 70, 79, 89).

3. Построить эмпирические функции распределения (абсолютные и накопленные частоты) успеваемости в группе из 20 студентов:

Вариант №	Данные
1	4, 4, 5, 3, 4, 5, 4, 5, 3, 5, 3, 3, 5, 4, 5, 4, 3, 5, 3, 5.
2	5, 4, 4, 4, 3, 5, 4, 4, 5, 5, 4, 3, 4, 5, 5, 4, 3, 3, 4, 5.
3	5, 5, 3, 4, 3, 3, 4, 4, 3, 4, 4, 3, 3, 3, 4, 3, 5, 5, 3, 5.

4	5, 5, 5, 5, 5, 3, 4, 5, 4, 4, 3, 4, 3, 4, 3, 3, 4, 3, 3, 3.
5	3, 3, 3, 4, 4, 4, 3, 4, 3, 4, 5, 4, 3, 5, 5, 3, 5, 3, 4, 3.
6	5, 5, 5, 4, 3, 3, 3, 4, 3, 4, 5, 4, 4, 5, 4, 4, 4, 4, 4, 4.
7	5, 3, 4, 5, 5, 4, 3, 3, 4, 3, 3, 3, 4, 3, 3, 4, 4, 5, 3, 4.
8	4, 4, 3, 5, 5, 3, 3, 3, 3, 3, 5, 4, 5, 3, 3, 5, 3, 5, 5, 4.
9	5, 5, 3, 5, 4, 5, 3, 3, 4, 3, 3, 5, 3, 3, 3, 3, 5, 5, 3, 5.
10	4, 3, 3, 3, 4, 4, 4, 3, 3, 4, 3, 4, 5, 5, 4, 5, 4, 4, 3, 5.

4. Найти среднее значение и стандартное отклонение результатов бега на дистанцию 100 м. у группы студентов:

Вариант №	Данные
1	12,8; 13,2; 13,0; 12,9; 13,5; 13,1.
2	13,5; 13,7; 12,3; 13,0; 12,5; 12,8.
3	12,8; 13,6; 13,9; 13,6; 12,1; 12,2.
4	12,8; 13,2; 13,4; 13,8; 13,7; 13,1.
5	13,8; 12,9; 12,7; 12,4; 12,6; 13,2.
6	12,3; 13,8; 13,5; 13,9; 13,4; 12,5.
7	13,2; 12,7; 12,1; 14,0; 13,6; 12,6.
8	12,6; 12,0; 12,7; 12,9; 13,9; 13,3.
9	13,1; 12,2; 13,3; 13,7; 13,7; 12,8.
10	12,5; 12,7; 12,4; 12,5; 13,2; 13,4.

5. Найти выборочные среднее, медиану, моду, дисперсию и стандартное отклонение для следующей выборки:

Вариант №	Данные
1	26, 35, 29, 27, 33, 35, 30, 33, 31, 29.
2	34, 29, 34, 26, 27, 33, 31, 35, 28, 25.
3	31, 32, 26, 29, 26, 25, 25, 30, 34, 31.
4	29, 33, 31, 33, 27, 27, 34, 30, 35, 30.
5	31, 26, 30, 26, 35, 35, 26, 28, 35, 25.
6	31, 30, 29, 34, 33, 32, 34, 31, 26, 27.
7	33, 26, 25, 28, 30, 25, 34, 26, 34, 25.
8	35, 25, 35, 33, 30, 29, 26, 35, 29, 27.
9	34, 33, 30, 33, 35, 35, 28, 33, 35, 30.
10	28, 27, 28, 33, 33, 29, 25, 27, 32, 34.

6. Построить функцию, наилучшим образом отражающую данную зависимость:

Вариант №	Данные					
1	x	1,0	1,5	3,0	4,5	5,0
	y	1,25	1,4	1,5	1,75	2,25
2	x	1,0	1,5	3,0	4,5	5,0
	y	1	1,25	1,45	1,7	1,95
3	x	1,0	1,5	3,0	4,5	5,0
	y	1,15	1,4	1,55	1,8	2,05

4	x	1,0	1,5	3,0	4,5	5,0
	y	1,05	1,3	1,55	1,8	1,95
5	x	1,0	1,5	3,0	4,5	5,0
	y	1	1,2	1,45	1,7	1,85
6	x	1,0	1,5	3,0	4,5	5,0
	y	1,35	1,5	1,75	1,85	2,15
7	x	1,0	1,5	3,0	4,5	5,0
	y	1,15	1,45	1,65	1,8	1,95
8	x	1,0	1,5	3,0	4,5	5,0
	y	1,05	1,25	1,45	1,6	1,8
9	x	1,0	1,5	3,0	4,5	5,0
	y	1	1,15	1,35	1,65	1,95
10	x	1,0	1,5	3,0	4,5	5,0
	y	1,2	1,45	1,6	1,85	2,1

7. В 80-е годы уровень дефицита бюджета в разных странах складывался следующим образом:

Вариант №	Данные								
1	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	A	2,9	2,3	3,1	2,2	2,0	2,7	6,5	8,0	9,1
	B	2,8	2,6	4,1	6,3	5,0	5,4	5,3	3,4	3,2
2	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	C	2,4	3,1	4,0	4,2	5,5	4,6	6,8	7,8	9,3
	D	2,6	2,9	3,8	4,9	5,9	5,5	4,2	4,0	3,5
3	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	E	2,2	2,8	3,4	4,5	5,3	5,1	7,3	7,0	6,4
	F	2,4	2,9	3,5	4,7	5,9	6,1	5,8	5,1	4,2
4	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	G	2,6	3,7	4,2	4,0	5,7	6,8	6,4	6,1	5,4
	H	2,7	3,1	4,2	5,3	5,5	6,2	6,9	6,3	5,4
5	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	I	2,6	3,4	4,0	4,7	5,3	5,2	5,9	5,4	4,8
	J	2,7	3,4	3,8	4,5	4,9	5,7	6,4	5,7	5,1
6	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	K	2,3	3	3,1	2,5	4	4,2	5,5	6	7,2
	L	2	2,5	3,8	4,7	6,2	5,7	5,2	5,5	6
7	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	M	3	4	4,3	5	5,4	6	6,1	7	8
	N	1,9	2,4	3,3	3,7	4,1	5	6,3	6,5	7,5
8	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	O	2,8	3,1	3,5	4	4,8	5	4,7	5,3	6,1
	P	2,5	3	3,6	4	4,2	4,5	5,1	6,4	8

9	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	Q	2,4	2,9	3,3	3,6	4,1	4,7	5	5,3	6,7
	R	1,8	2	2,1	2,3	2,5	2,9	4	5,5	7
10	Страна	Годы								
		1980	1981	1982	1983	1984	1985	1986	1987	1988
	S	3,1	3,3	4	4,3	4,2	5	5,2	6	7,1
	T	2	2,3	2,6	4,2	4,3	4	4,4	5	6

Построить функции, наилучшим образом отражающие зависимости дефицита бюджета от времени в обеих странах.

8. Количество вложенных в производство средств и полученная в результате прибыль соотносятся следующим образом:

Вариант №	Данные						
1	x	1,6	2,0	2,5	3,0	4,0	7,0
	y	8,5	9,0	11,0	13,0	22,0	70,0
2	x	1,5	2,0	3,0	4,5	5,0	7,3
	y	9,2	11,0	15,3	18,6	25,0	78,5
3	x	1,0	2,5	3,5	5,5	6,0	8,9
	y	8,4	10,3	14,2	19,6	28,7	73,2
4	x	1,8	2,8	4,0	5,2	7,0	9,3
	y	9,6	12,5	15,7	21,2	28,7	66,6
5	x	1,6	2,2	3,1	4,0	4,9	6,6
	y	9,2	12,0	13,0	20,5	31,0	65,0
6	x	1,8	2,7	4,2	5,5	7,0	9,5
	y	7,4	11,2	16,0	19,3	23,5	71,6
7	x	1,6	2,8	3,7	4,2	5,9	7,8
	y	8,2	12,3	15,6	17,5	22,2	74,4
8	x	1,0	2,5	3,5	4,0	5,2	8,1
	y	7,0	10,3	18,2	25	40	101
9	x	1,5	2,5	3,1	4,5	5,4	8,2
	y	10	12	13,3	15,5	36,6	69
10	x	1,1	1,5	3,2	3,5	4,5	7,5
	y	8,1	11,1	14,7	20,0	31,1	77

Записать аналитическую зависимость между x и y. Проанализировать полученный ответ. Каковы перспективы предприятия? Какая будет прибыль, если вложить 10.0 единиц? Сколько надо вложить средств, чтобы получить прибыль 100.0 единиц?

9. Застройщик оценивает стоимость группы небольших офисных зданий в традиционном деловом районе. Оценку цены офисного здания в заданном районе застройщик предполагает осуществлять на основе следующих переменных: y - оценочная цена здания под офис, x1- общая площадь в квадратных метрах, x1- количество офисов, x3 - количество входов, x, - время эксплуатации здания в годах. Предполагается, что существует линейная зависимость между каждой независимой переменной (x1, x2, x3 и x4) и зависимой переменной (y), то есть ценой здания под офис в данном районе. Застройщик наугад выбирает 11 зданий из имеющихся 1500 и получает следующие данные (общие для всех):

X1	X2	X3	X4	Y
2310	2	2	20	14200
2333	2	2	12	144000
2356	3	1,5	33	151000
2379	3	2	43	150000
2402	2	3	53	139000

2425	4	2	23	169000
2448	2	1,5	99	126000
2471	2	2	34	142900
2494	3	3	23	163000
2517	4	4	55	169000
2540	2	3	22	149000

Здесь «полвхода» (1/2) означает вход только для доставки корреспонденции. Найти параметры аппроксимирующего уравнения. С помощью функции **ТЕНДЕНЦИЯ** определить оценочную стоимость здания под офис в том же районе, которое имеет площадь 2500 квадратных метров, три офиса, два входа, зданию 25 лет.

10. Найти наиболее популярный туристический маршрут из четырех реализуемых фирмой (моду), если за неделю последовательно были реализованы следующие маршруты (приводятся номера маршрутов):

Вариант №	Данные
1	1, 3, 3, 2, 1, 1, 4, 4, 2, 4, 1, 3, 2, 4, 1, 4, 4, 3, 1, 2, 3, 4, 1, 1, 3.
2	2, 2, 4, 2, 2, 4, 2, 3, 1, 2, 2, 4, 4, 3, 1, 1, 3, 2, 4, 2, 2, 2, 2, 3, 2.
3	1, 3, 1, 3, 1, 2, 2, 2, 3, 2, 2, 3, 1, 1, 2, 4, 3, 2, 4, 4, 1, 2, 1, 3, 2.
4	1, 4, 4, 2, 4, 4, 4, 2, 1, 1, 1, 2, 1, 4, 1, 1, 3, 2, 2, 3, 2, 3, 1, 3, 2.
5	3, 4, 2, 4, 1, 4, 3, 2, 1, 1, 2, 2, 1, 3, 2, 3, 2, 4, 4, 2, 3, 3, 2, 4, 3.
6	3, 4, 4, 4, 4, 3, 3, 2, 4, 2, 1, 2, 2, 1, 2, 3, 4, 4, 4, 3, 3, 3, 4, 4, 1.
7	1, 3, 2, 1, 4, 1, 2, 3, 1, 2, 4, 3, 3, 2, 2, 3, 1, 1, 2, 2, 1, 4, 4, 2, 3.
8	3, 3, 2, 1, 1, 1, 1, 4, 4, 3, 1, 2, 1, 4, 1, 4, 4, 1, 4, 3, 4, 2, 3, 3, 4.
9	2, 2, 3, 3, 4, 1, 4, 4, 3, 1, 3, 4, 2, 1, 2, 4, 4, 3, 3, 2, 2, 1, 1, 1, 3.
10	2, 4, 4, 4, 4, 2, 4, 4, 2, 3, 3, 2, 2, 1, 3, 1, 2, 3, 4, 3, 4, 2, 1, 2, 3.

11. В рабочей зоне производились замеры концентрации вредного вещества. Получен ряд значений (в мг/м3):

Вариант №	Данные
1	12, 16, 15, 14, 10, 20, 16, 14, 18, 14, 15, 17, 23, 16.
2	14, 21, 13, 15, 22, 21, 21, 10, 14, 13, 17, 13, 24, 24.
3	24, 10, 17, 24, 21, 21, 25, 21, 20, 17, 14, 20, 20, 20.
4	10, 10, 17, 17, 18, 20, 19, 12, 21, 23, 25, 11, 15, 25.
5	12, 13, 14, 19, 13, 18, 17, 24, 20, 14, 11, 13, 13, 18.
6	10, 22, 11, 11, 11, 10, 10, 14, 22, 10, 21, 13, 15, 21.
7	15, 20, 24, 19, 16, 25, 18, 17, 12, 20, 16, 11, 13, 11.
8	16, 12, 16, 15, 17, 10, 11, 25, 21, 24, 10, 14, 25, 24.
9	16, 14, 24, 10, 13, 21, 19, 25, 15, 17, 22, 13, 25, 19.
10	19, 17, 20, 25, 23, 24, 23, 20, 20, 15, 21, 14, 20, 15.

Необходимо определить основные выборочные характеристики.

12. Определить, лежит ли значение X внутри границ 95-процентного доверительного интервала выборки:

Вариант №	X	Данные
1	19	2, 3, 5, 7, 4, 9, 6, 4, 9, 10, 4, 7, 19.
2	13	9, 4, 9, 10, 4, 7, 8, 4, 4, 5, 10, 6, 13.
3	15	7, 1, 4, 10, 9, 7, 3, 3, 10, 3, 10, 10, 15.

4	14	9, 4, 5, 1, 7, 2, 8, 5, 6, 2, 1, 5, 14.
5	15	9, 3, 3, 1, 9, 5, 6, 5, 6, 8, 9, 7, 15.
6	11	3, 8, 2, 1, 6, 9, 8, 6, 5, 10, 1, 1, 11.
7	18	2, 8, 7, 3, 7, 6, 1, 5, 10, 10, 2, 6, 18.
8	16	5, 3, 4, 6, 8, 6, 4, 2, 5, 9, 8, 10, 16.
9	13	1, 8, 10, 7, 5, 5, 1, 9, 3, 2, 9, 2, 13.
10	10	8, 4, 9, 1, 8, 4, 7, 4, 6, 6, 6, 4, 10.

13. Определите с уровнем значимости a = 0,05 максимальное отклонение среднего значения генеральной совокупности от среднего выборки:

Вариант №	Данные
1	3, 4, 4, 2, 5, 3, 4, 3, 5, 4, 3, 5, 6.
2	6, 4, 4, 6, 2, 4, 6, 3, 6, 4, 2, 5, 2.
3	5, 4, 4, 5, 3, 3, 6, 5, 6, 2, 2, 2, 5.
4	4, 2, 4, 4, 2, 2, 6, 4, 2, 6, 4, 5, 6.
5	3, 4, 6, 6, 3, 2, 6, 6, 5, 5, 5, 6, 6.
6	3, 2, 5, 2, 4, 4, 5, 4, 6, 5, 6, 3, 3.
7	6, 5, 5, 4, 4, 2, 2, 4, 6, 3, 5, 2, 5.
8	6, 3, 5, 5, 4, 5, 5, 3, 6, 6, 5, 2, 4.
9	2, 6, 4, 4, 5, 2, 4, 6, 2, 2, 2, 3, 3.
10	5, 6, 3, 4, 6, 2, 6, 5, 4, 4, 5, 2, 3.

14. Найти соответствие экспериментальных данных нормальному закону распределения для следующей выборки весов детей (кг):

Вариант №	Данные
1	21, 21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 27, 27.
2	30, 30, 30, 30, 30, 30, 30, 30, 30, 31, 31, 31, 31, 31, 31, 31, 32, 32, 32, 32, 32, 32, 33, 33, 33, 33, 33, 33, 33, 33, 33, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 35, 35, 35, 35, 35, 35, 35, 35, 35, 36, 36, 36, 36, 36, 36, 36, 36, 36.
3	15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 17, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22, 22.
4	10, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12, 12, 12, 13, 13, 13, 13, 13, 13, 13, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16.
5	23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 27, 27, 27, 27, 27, 27, 27, 27, 27, 28, 28, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29, 29, 29.
6	20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26, 26, 26.
7	5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12, 12.

8	21, 21, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27.
9	17, 17, 17, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 24, 24.
10	26, 26, 26, 26, 26, 26, 26, 27, 27, 27, 27, 27, 27, 27, 27, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29, 29, 30, 30, 30, 30, 30, 30, 30, 30, 30, 31, 31, 31, 31, 31, 31, 31, 31, 31, 32, 32, 32, 32, 32, 32, 32, 32, 33, 33.

15. Даны результаты бега на дистанцию 100 м в секундах в двух группах студентов. Студенты первой группы в течение года посещали факультативные занятия по физкультуре. Определить, достоверны ли отличия по результатам бега в этих группах.

Вариант №	Данные	
	Посещавшие факультатив	Не посещавшие
1	12,6, 12,3, 11,9, 12,2, 13,0, 12,4.	12,8, 13,2, 13,0, 12,9, 13,5, 13,1.
2	12,9, 12,4, 12,3, 12,2, 12,1, 12,1.	13,4, 12,8, 12,1, 13,3, 12,2, 12,9.
3	12,4, 12,4, 12,3, 12,1, 12,3, 12,2.	12,3, 13,0, 12,1, 12,3, 12,0, 12,2.
4	12,7, 12,5, 12,6, 12,8, 13,0, 12,4.	12,9, 13,4, 13,2, 13,2, 12,2, 13,3.
5	12,6, 12,4, 12,5, 12,3, 12,4, 12,5.	12,9, 12,9, 13,1, 13,3, 13,2, 13,3.
6	12,1, 12,5, 12,9, 12,7, 12,4, 12,1.	12,7, 12,8, 13,4, 12,3, 13,1, 12,3.
7	12,1, 12,1, 12,6, 12,7, 12,7, 12,8.	13,4, 12,6, 13,1, 13,4, 13,2, 12,2.
8	12,6, 12,1, 12,8, 12,9, 12,6, 12,7.	13,1, 13,2, 13,5, 12,3, 13,1, 13,4.
9	12,5, 12,5, 13,0, 12,6, 12,6, 12,7.	12,9, 12,5, 12,7, 13,3, 13,5, 13,0.
10	12,0, 12,2, 12,2, 12,7, 12,4, 12,7.	13,3, 12,4, 12,1, 12,5, 12,1, 13,2.

16. В ходе социологического опроса на вопрос о перенесенном в детстве заболевании ответы распределились следующим образом:

Вариант №	Данные			
1		Да	Нет	Не помню
	Мужчины	58	11	10
	Женщины	35	25	23
2		Да	Нет	Не помню
	Мужчины	47	18	13
	Женщины	42	12	31
3		Да	Нет	Не помню
	Мужчины	45	17	25
	Женщины	30	28	18
4		Да	Нет	Не помню
	Мужчины	55	19	8
	Женщины	38	14	29
5		Да	Нет	Не помню
	Мужчины	47	21	12
	Женщины	40	18	16
6		Да	Нет	Не помню
	Мужчины	36	29	14
	Женщины	45	20	11
7		Да	Нет	Не помню
	Мужчины	52	22	12
	Женщины	38	29	16
8		Да	Нет	Не помню
	Мужчины	42	20	21
	Женщины	28	29	17
9		Да	Нет	Не помню
	Мужчины	36	25	13
	Женщины	27	19	21
10		Да	Нет	Не помню
	Мужчины	48	15	19
	Женщины	35	22	31

Есть ли достоверные отличия в ответах женщин и мужчин?

17. Приведены данные ежемесячной результативности (количество голов) футбольной команды в двух сезонах:

Вариант №	Данные									
	Месяцы	3	4	5	6	7	8	9	10	11
1	2008 г.	3	4	5	8	9	1	2	4	5
	2009 г.	6	19	3	2	14	4	5	17	1
2	2008 г.	16	12	1	7	12	7	1	10	19
	2009 г.	6	10	8	14	10	8	2	15	13
3	2008 г.	8	10	16	17	16	5	2	5	10
	2009 г.	8	7	2	12	14	4	3	9	15
4	2008 г.	11	5	13	3	15	18	19	4	4
	2009 г.	15	9	15	13	13	16	4	3	5
5	2008 г.	18	19	11	4	3	14	16	11	9
	2009 г.	8	6	12	16	13	2	8	14	8
6	2008 г.	9	1	7	2	11	5	12	6	19

	2009 г.	18	19	7	10	15	17	19	3	10
7	2008 г.	14	7	15	7	19	17	8	8	16
	2009 г.	7	3	5	9	10	10	18	1	10
8	2008 г.	8	12	19	9	8	6	8	4	11
	2009 г.	5	18	16	6	4	4	3	13	15
9	2008 г.	9	10	14	18	18	10	1	17	18
	2009 г.	9	2	9	16	14	4	7	12	5
10	2008 г.	4	5	19	6	14	16	18	6	5
	2009 г.	17	4	16	11	11	10	9	16	2

Определить, есть ли статистические различия в ежемесячной результативности команды в рассматриваемых сезонах?

18. Определить, достоверны ли различия в количестве приобретаемых туристических путевок семейными парами и отдельными туристами.

Вариант №	Данные						
	Месяцы	1	2	3	4	5	6
1	Пары	67	75	58	89	96	94
	Одиночки	43	56	78	87	85	90
2	Пары	51	78	63	43	68	51
	Одиночки	72	66	70	41	58	91
3	Пары	41	54	71	40	74	81
	Одиночки	61	50	59	73	57	96
4	Пары	40	72	93	90	86	83
	Одиночки	49	69	98	76	65	87
5	Пары	55	71	65	77	76	51
	Одиночки	78	65	47	45	65	66
6	Пары	50	72	93	86	69	43
	Одиночки	58	49	46	100	100	72
7	Пары	42	92	96	43	66	95
	Одиночки	85	51	72	61	48	71
8	Пары	75	54	84	52	70	96
	Одиночки	49	82	83	64	80	96
9	Пары	100	73	82	76	74	56
	Одиночки	90	50	75	88	54	59
10	Пары	93	73	47	67	47	46
	Одиночки	100	52	58	62	72	89

19. В таблице приведены результаты группы студентов по скоростному чтению до и после специального курса по быстрому чтению.

Вариант №	Данные									
	Студент	1	2	3	4	5	6	7	8	9	10
1	До курса	86	83	86	70	66	90	70	85	77	86
	После	82	79	91	77	68	86	81	90	85	94
2	До курса	68	100	78	94	89	61	72	78	64	76
	После	95	94	82	92	82	97	68	90	99	60
3	До курса	72	93	67	99	75	64	70	83	98	77
	После	78	77	77	82	85	99	63	72	87	70
4	До курса	63	96	78	78	82	99	99	92	64	80
	После	79	90	67	68	96	63	77	60	99	74
5	До курса	75	100	95	79	83	94	94	83	65	93
	После	81	95	60	74	61	82	100	85	61	86

6	До курса	73	60	67	74	74	73	99	85	76	97
	После	74	60	99	73	79	99	93	68	95	77
7	До курса	93	66	80	87	76	76	65	63	91	98
	После	100	100	81	69	82	66	78	82	85	67
8	До курса	98	87	85	82	68	76	67	95	84	99
	После	71	84	85	73	69	94	92	99	86	64
9	До курса	91	77	92	64	62	87	83	80	64	74
	После	76	84	77	91	90	69	90	62	85	60
10	До курса	64	66	82	99	91	89	90	78	92	86
	После	96	60	88	82	74	100	62	82	100	92

Произошли ли статистически значимые изменения скорости чтения у студентов?

20. Определить, влияет ли фактор образования на уровень зарплаты сотрудников в гостинице на основании следующих данных:

Вариант №	Данные						
	Образование	Зарплата сотрудника					
1	Высшее	3200	3000	2600	2000	1900	1900
	Среднее спец.	2600	2000	2000	1900	1800	1700
	Среднее	2000	2000	1900	1800	1700	1700
2	Высшее	2500	2700	2000	2100	2900	2600
	Среднее спец.	2000	2400	2100	1900	2100	1800
	Среднее	2300	1900	1800	1700	2000	1900
3	Высшее	2100	3300	2700	2600	2300	2500
	Среднее спец.	2000	2200	1900	2500	2300	2000
	Среднее	2200	1900	1700	2000	2100	1800
4	Высшее	2500	2600	2300	2500	2800	2400
	Среднее спец.	2400	2000	2300	1900	1700	1800
	Среднее	1700	1900	2000	1800	2100	1900
5	Высшее	2600	2300	2700	2500	2200	2400
	Среднее спец.	2300	2000	2200	1900	2400	2300
	Среднее	1900	1700	2100	1800	1800	1900
6	Высшее	2400	2600	2300	2400	2100	2300
	Среднее спец.	2200	1900	2300	2400	2000	2100
	Среднее	1700	1900	2100	2000	1800	1800
7	Высшее	2600	2400	2300	2500	2700	2500
	Среднее спец.	2100	2300	2500	2200	2400	2000
	Среднее	1900	2100	1800	2000	1700	1800
8	Высшее	2700	2900	2500	3200	2600	3000
	Среднее спец.	2400	2200	2700	2300	2500	2000
	Среднее	1900	1600	1800	2000	1900	2200
9	Высшее	2600	2700	3100	2800	2500	2900
	Среднее спец.	2400	2100	2600	2300	2100	2500
	Среднее	2300	2000	2100	1700	1900	1800
10	Высшее	3400	3000	2800	2500	2700	3100
	Среднее спец.	2400	2700	2200	2300	2500	2100
	Среднее	1900	1700	2000	2100	1800	1900

21. Определить, имеется ли взаимосвязь между рождаемостью и смертностью (количество на 1000 человек) в Санкт-Петербурге:

Вариант №	Данные		
	Годы	Рождаемость	Смертность
1	1991	9,3	12,5
	1992	7,4	13,5
	1993	6,6	17,4
	1994	7,1	17,2
	1995	7,0	15,9
	1996	6,6	14,2
2	1991	8,0	16,6
	1992	6,9	16,1
	1993	7,0	13,6
	1994	8,8	14,2
	1995	7,0	17,5
	1996	9,8	11,0
3	1991	7,3	17,8
	1992	9,8	17,6
	1993	8,8	15,3
	1994	7,5	12,4
	1995	8,2	14,1
	1996	6,8	14,7
4	1991	8,1	16,1
	1992	8,4	13,2
	1993	6,9	13,3
	1994	7,0	11,7
	1995	6,7	17,1
	1996	9,4	13,8
5	1991	9,7	16,6
	1992	9,2	14,6
	1993	7,2	12,8
	1994	7,9	15,6
	1995	9,4	15,9
	1996	6,3	12,3
6	1991	10,0	12,6
	1992	8,3	13,0
	1993	8,5	15,7
	1994	7,5	15,8
	1995	6,6	14,7
	1996	6,3	15,2
7	1991	6,0	14,2
	1992	6,8	14,7
	1993	9,1	12,7
	1994	6,1	16,4
	1995	10,0	17,3
	1996	6,4	12,5

8	1991	8,1	12,3
	1992	8,6	15,1
	1993	9,2	13,6
	1994	7,6	15,5
	1995	8,2	15,2
	1996	6,9	15,5
9	1991	7,2	17,3
	1992	6,9	12,9
	1993	7,4	13,4
	1994	8,0	16,8
	1995	9,2	17,0
	1996	6,4	17,7
10	1991	6,7	14,2
	1992	10,0	17,4
	1993	6,6	14,3
	1994	9,3	16,7
	1995	8,3	14,3
	1996	6,7	12,2

22. Определить, имеется ли взаимосвязь между годовым уровнем инфляции (%), ставкой рефинансирования (%) и курсом доллара (р./$), по следующим данным ежегодных наблюдений:

Вариант №	Данные		
	Уровень инфляции	Ставка рефинансирования	Курс $
1	84	85	6,3
	45	55	14
	56	65	20
	34	40	28
	23	28	29
2	80	77	7
	53	50	13
	61	62	21
	38	42	29
	25	26	33
3	79	84	8,2
	47	53	16
	60	67	22
	36	44	29
	27	33	35
4	92	90	7,6
	66	72	12
	78	84	16
	45	47	21
	34	39	29

5	77	78	8,1
	49	52	13
	57	63	17
	38	41	21
	26	29	26
6	84	86	7,9
	53	57	11
	67	70	18
	42	45	22
	30	36	28
7	77	81	8,5
	50	53	12
	63	68	16
	42	47	21
	31	34	26
8	86	90	9,2
	54	62	12
	69	78	18
	42	45	23
	34	37	29
9	80	84	7,6
	52	56	10
	66	73	16
	42	43	20
	31	28	25
10	79	82	9,8
	58	63	12
	63	68	18
	47	55	22
	34	36	27

23. Построить зависимость зарплаты (р.) от возраста сотрудника гостиницы по следующим данным:

Вариант №	Данные						
	Возраст	20	50	45	40	25	30
1	Зарплата	800	2500	2500	2000	1200	1800
2	Зарплата	900	2600	2500	2200	1300	1900
3	Зарплата	700	2700	2500	2300	1100	1700
4	Зарплата	1000	2400	2300	1900	1200	1500
5	Зарплата	900	2400	2400	2000	1400	1700
6	Зарплата	1000	2500	2300	2100	1200	1900
7	Зарплата	700	2500	2400	1900	1300	1600
8	Зарплата	900	2300	2300	1900	1200	1800
9	Зарплата	800	2600	2600	2100	1300	1700
10	Зарплата	800	2600	2500	2200	1200	1800

24. Построить зависимость жизненной емкости легких в литрах (Y) от роста в метрах (X_1) и возраста в годах (X_2) для группы мужчин:

Вариант №	Данные		
	X_1	X_2	Y
1	1,85	18	5,4
	1,8	25	6,7
	1,75	20	4,8
	1,7	24	5,1
	1,68	21	4,5
	1,73	19	4,8
	1,77	22	5,11
	1,81	23	5,6
	1,76	18	4,7
2	1,85	20	6,2
	1,8	18	4,7
	1,75	18	5,4
	1,7	21	5,6
	1,68	25	4,7
	1,73	21	4,7
	1,77	24	5,8
	1,81	25	5
	1,76	17	5,8
3	1,85	22	5,1
	1,8	18	5,7
	1,75	23	5,3
	1,7	20	6
	1,68	20	5,4
	1,73	23	6,3
	1,77	19	5,7
	1,81	22	5,7
	1,76	21	4,7
4	1,85	19	5,2
	1,8	18	5,7
	1,75	18	4,7
	1,7	21	5,3
	1,68	21	6,3
	1,73	23	4,7
	1,77	24	5,1
	1,81	17	4,9
	1,76	20	5,3
5	1,85	21	5,4
	1,8	21	6,1
	1,75	24	5
	1,7	25	6
	1,68	18	5,5
	1,73	23	5
	1,77	22	5,4
	1,81	23	6,2
	1,76	18	6
6	1,85	19	6,2
	1,8	18	6
	1,75	22	6,1

	1,7	19	5,6
	1,68	25	5,7
	1,73	22	5,5
	1,77	17	5,4
	1,81	21	4,9
	1,76	20	4,6
7	1,85	17	4,8
	1,8	23	4,5
	1,75	23	5,6
	1,7	21	5,6
	1,68	18	4,9
	1,73	17	5,1
	1,77	24	5,4
	1,81	19	5,6
	1,76	17	4,5
8	1,85	18	6,2
	1,8	24	5,5
	1,75	23	4,7
	1,7	20	5,2
	1,68	22	4,9
	1,73	25	6,1
	1,77	18	5,8
	1,81	19	4,8
	1,76	20	5,2
9	1,85	19	5,2
	1,8	23	5,5
	1,75	25	4,7
	1,7	25	4,8
	1,68	23	5,1
	1,73	23	5,5
	1,77	21	4,7
	1,81	25	5,8
	1,76	21	6
10	1,85	25	4,9
	1,8	25	6
	1,75	19	4,6
	1,7	22	5,4
	1,68	21	6,2
	1,73	25	5,9
	1,77	20	4,6
	1,81	24	4,8
	1,76	20	4,8

25. Определить должное значение жизненной емкости легких для мужчины возраста 22-х лет и роста 183 см из регрессионного уравнения, полученного в предыдущем упражнении.

26. Имеются данные о цене на нефть x (ден. ед.) и индексе акций нефтяных компаний y (усл. ед.):

Вариант №	Данные						
1	x	17,28	17,05	18,30	18,80	19,20	18,50
	y	537	534	550	555	560	552

2	x	18,71	19,44	18,65	17,18	17,77	17,39
	y	556,3	555,94	544,9	540,72	541,07	559,47
3	x	18,99	17,71	18,82	17,42	18,4	18,15
	y	549,66	555,85	541,18	544,68	537,08	538,97
4	x	19,11	18,72	18,56	18,99	18,77	18,77
	y	559,48	552,02	533,85	539,22	549,42	553,12
5	x	17,86	19,29	17,58	18,83	17,75	17,49
	y	537,66	535,68	541,28	546,18	550,84	551,58
6	x	18,42	18,15	18,97	17,61	17,91	18,96
	y	557,44	535,81	552,21	540,84	536,83	548,6
7	x	18,33	17,2	17,39	18,97	17,94	18,59
	y	555,61	544,63	549,14	558,76	554,71	549,29
8	x	17,57	19,31	18,4	17,44	17,58	17,12
	y	557,69	539,13	559,75	557,16	536,89	537,44
9	x	19,3	18,84	17,92	17,87	17,98	18,13
	y	530,91	559,49	550,67	548,35	536,02	536,65
10	x	17,11	18,17	19,04	18,85	18,71	17,52
	y	532,44	548,24	549,15	534,32	530,48	543,26

Построить зависимость индекса акций нефтяных компаний от цены на нефть.