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 форм обучения

Могилев 2024

УДК 519.6
ББК 22.176
С48

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Автоматизированные системы управления» «24» сентября 2024 г., протокол № 2

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

Учебное издание

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Ответственный за выпуск	А. И. Якимов
Корректор	И. В. Голубцова
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 21 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 «Статистические методы обработки данных» является обучение студентов основным 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 потребуются:

– язык R (https://cran.r-project.org/bin/windows/base/);

– среда разработки RStudio (https://www.rstudio.com/products/rstudio/);

– учебные материалы с сайта http://www.qsar4u.com/files/rintro/01.html.

Пакет MS Exce1. Пакет MS Exce1 оснащен средствами статистической обработки данных. И хотя Exce1 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 несплошное. 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 несплошного 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, т. е. 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, т. к. 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.

Практические задания

1 Имеются данные о цене на нефть x (д. е.) и индексе акций нефтяных компаний y (у. е.) (таблица 1.1).

Таблица 1.1 – Данные о цене на нефть и индексе акций нефтяных компаний

Номер варианта	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

Окончание таблицы 1.1

Номер варианта	Данные						
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 таблицы 1.2.

Таблица 1.2 – Данные о зарплате и возрасте сотрудников

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

Контрольные вопросы

1 Какие пакеты в R можно использовать для работы с описательной статистикой?

2 Что такое среднее значение и как его можно вычислить средствами R?

3 Какую функцию использовать для расчета медианы в R?

4 Каким образом можно найти минимальное и максимальное значения переменной в R?

5 Какая функция позволяет найти моду в R?
6 Что такое стандартное отклонение и как его вычислить в R?
7 Как посчитать коэффициент вариации средствами R?
8 Как проверить наличие выбросов в данных с помощью R?
9 Как вычислить квантиль заданного уровня в R?
10 Как построить гистограмму распределения данных средствами R?

2 Лабораторная работа № 2. Графические возможности среды программирования R для визуализации 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 2.1. В этом наборе данных 50 наблюдений и две переменных (скорость, миль/ч, и длина тормозного пути в футах).

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

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

Теперь *d* имеет тип данных *data.frame* (набор данных), в чем можно удостовериться, посмотрев в правом верхнем углу окна таблицу среды *Environment* (рисунок 2.2).

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,...
```

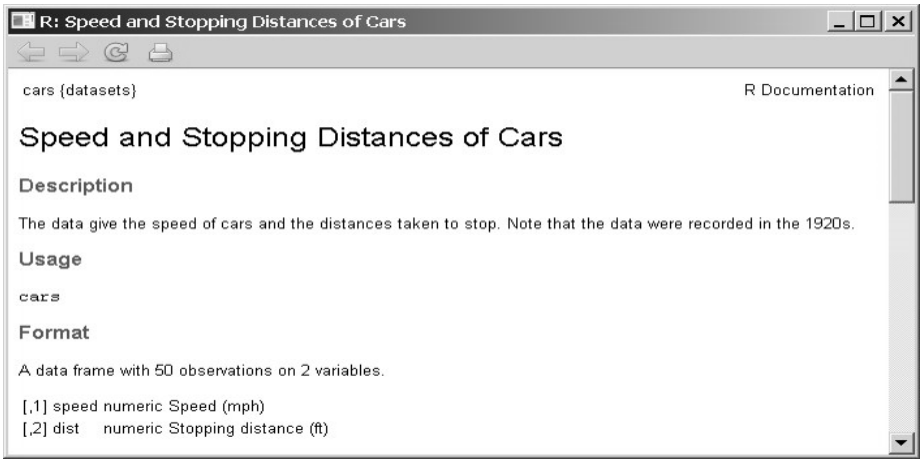


Рисунок 2.1 – Справка по набору данных `cars`

Рисунок 2.2 – Описание набора данных *d*

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 gplot, 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 2.3.

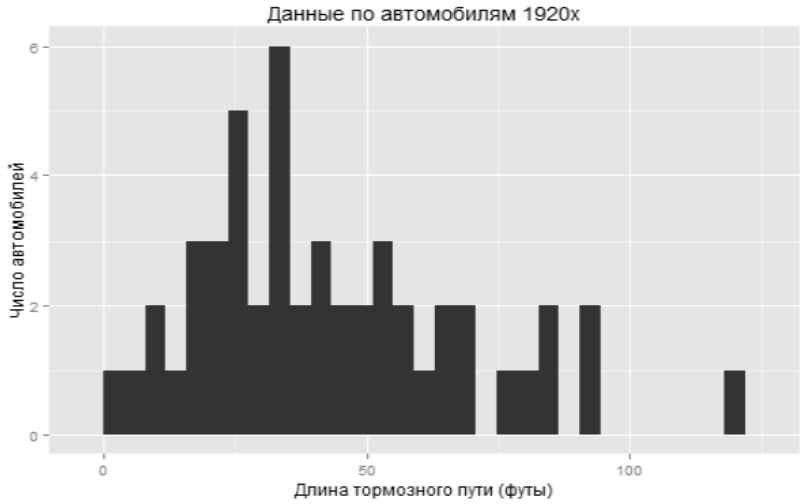


Рисунок 2.3 – Гистограмма абсолютных частот для переменной dist

Пример – Имеются данные о цене на нефть *x* (д. е.) и индексе акций нефтяных компаний *y* (у. е.) (таблица 2.1).

Таблица 2.1 – Данные о цене на нефть и индексе акций нефтяных компаний

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 графическая информация.

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

Результат выполнения функции pairs на рисунке 2.4.

В таблице (см. рисунок 2.4) два столбца. График на пересечении oil_price и action_price (слева внизу) показывает зависимость цены акций от цены на нефть.

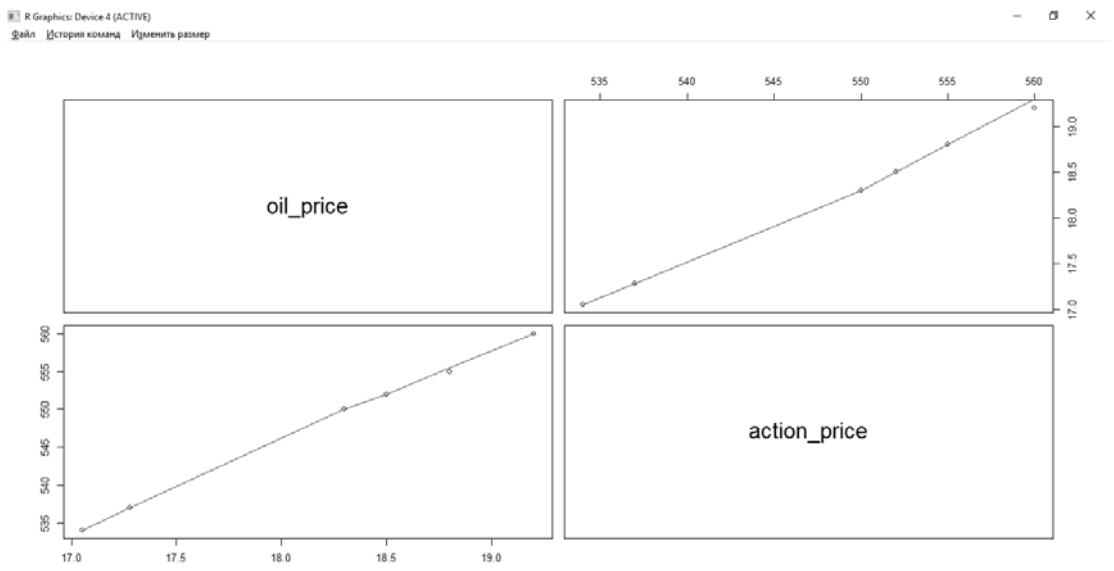


Рисунок 2.4 – Итоговые таблицы после выполнения функции pairs

Практические задания

1 Имеются данные о цене на нефть *x* (д. е.) и индексе акций нефтяных компаний *y* (у. е.) (см. таблицу 2.1).

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

2 Построить зависимость зарплаты (р.) от возраста сотрудника гостиницы по данным таблицы 2.2 средствами Excel и R.

Контрольные вопросы

1 Какие функции в среде программирования R используются для создания графиков, отображающих распределение числовых переменных?

2 Как можно визуализировать связь между двумя числовыми переменными с помощью графиков в R?

3 Какие типы диаграмм подходят для визуализации описательной статистики категориальных переменных в R?

4 Каким образом можно добавлять заголовки, оси координат, легенды и другие детали к графикам в R?

5 Как создать график boxplot в R для визуализации распределения числовой переменной?

6 Какие возможности предоставляет R для создания множественных графиков, отображающих сравнение нескольких групп данных?

7 Как можно изменить цвета, шрифты и другие атрибуты графика в R, чтобы сделать его более наглядным и привлекательным?

8 Какие типы графиков в R подходят для визуализации временных рядов или временных зависимостей данных?

9 Как создать график сопряженности для визуализации связей между несколькими переменными в R?

10 Как можно сохранить графики, созданные в среде R, в различных форматах (например, PNG, PDF, или JPEG) для последующего использования или публикации?

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 и сравнение с результатами в MS Excel.

Основные теоретические положения

Принятие статистических решений с помощью параметрических и непараметрических критериев средствами среды программирования R является важным аспектом в области статистического анализа данных. В данном контексте параметрические критерии используются для проверки гипотез о параметрах распределения данных, предполагая определенную форму распределения, например нормальное. Непараметрические критерии, в свою очередь, не требуют таких предположений о распределении и используются для оценки различий в данных, основываясь на их ранжировании.

В среде программирования R статистические решения с использованием параметрических критериев получают с помощью множества пакетов, таких как stats, car, lme4 и др. Эти пакеты предоставляют функции для проведения t-тестов, анализа дисперсии (ANOVA), регрессионного анализа и других методов, основанных на параметрических критериях.

Среда R также предоставляет широкий выбор пакетов для проведения непараметрических тестов, например пакет nparLD для проведения непараметрических тестов на различие медиан и пакет coin для применения непараметрических тестов в контексте рандомизированных исследований.

При использовании среды программирования R для принятия статистических решений с помощью параметрических и непараметрических критериев необходимо учитывать соответствие выбранного метода задаче и правильность интерпретации результатов. Также важно учитывать особенности данных, на которых проводятся тесты, и выбор наиболее подходящих методов в зависимости от предположений о распределениях и целях исследования.

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), т. е. 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 ч, а стандартное отклонение – 2,5 ч.

Решение

1 Откройте новую рабочую таблицу. Установите табличный курсор в ячейку А1.

2 Для определения границ доверительного интервала необходимо на панели 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 3.1). Нажмите кнопку *ОК*.

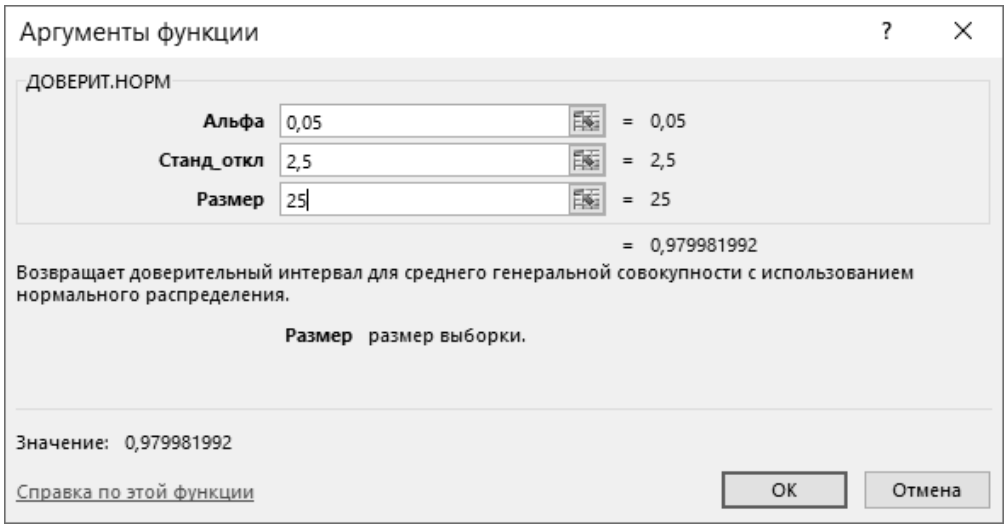


Рисунок 3.1 – Пример заполнения диалогового окна ДОВЕРИТ.НОРМ

4 В ячейке А1 появится полуширина 95-процентного 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) ч или от 139,02 до 140,98 ч.

Практические задания

1 Определить, лежит ли значение X внутри границ 95-процентного доверительного интервала выборки, представленной в таблице 3.1.

2 Определить с уровнем значимости $\alpha = 0{,}05$ максимальное отклонение среднего значения генеральной совокупности от среднего выборки (таблица 3.2).

3 Найти соответствие экспериментальных данных нормальному закону распределения для следующей выборки весов детей (кг) (таблица 3.3).

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 (таблица 3.4).

Таблица 3.1 – Выборка с исходными данными

Номер варианта	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

Таблица 3.2 – Выборка с исходными данными для практического задания

Номер варианта	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

Таблица 3.3 – Выборка с данными весов детей

Номер варианта	Данные
1	2
1	21, 21, 22, 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, 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, 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

Окончание таблицы 3.3

1	2
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, 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, 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, 22, 23, 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, 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

Таблица 3.4 – Выборка с результатами бега на дистанцию 100 м

Номер варианта	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

Контрольные вопросы

1 Что такое параметрический критерий и как он применяется для принятия статистических решений в R?

2 Как выбрать подходящий параметрический критерий в R для проверки гипотезы о различии средних?

3 Как применить непараметрический критерий в R для проверки гипотезы о различии медиан?

4 Какие данные требуются для применения параметрического критерия t-статистики в R?

5 Какие данные требуются для применения непараметрического критерия U-Манна-Уитни в R?

6 Какие данные требуются для применения параметрического критерия однофакторного дисперсионного анализа (ANOVA) в R?

7 Каким образом можно проверить нормальность распределения в R перед применением параметрического критерия?

8 Как применять параметрический критерий t-статистики в R для сравнения двух зависимых выборок?

9 Как применить непараметрический критерий знаковых ранговых разностей в R для сравнения двух зависимых выборок?

10 Как применить непараметрический критерий Краскела – Уоллиса в R для сравнения нескольких независимых выборок?

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 R проведение корреляционного анализа осуществляется с помощью мощных библиотек и функций, предназначенных специально для работы с данными. Для начала анализа необходимо загрузить данные в R и преобразовать их в нужный формат. Самая популярная библиотека для работы с данными в R – это tidyverse, которая содержит множество функций для удобной работы с таблицами и данными.

После загрузки данных можно использовать функции корреляционного анализа, такие как cor() для вычисления матрицы корреляций. Эта функция позволяет рассчитать корреляцию между всеми парами переменных в данных.

Для визуализации корреляционной матрицы можно использовать corrplot – библиотеку для построения различных типов графиков корреляций.

Также следует упомянуть о возможности вычисления коэффициентов корреляции с помощью функций cor.test(), которые позволяют проводить статистические тесты на значимость корреляций.

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

КОРРЕЛ(массив1: массив2),

где массив1 – диапазон ячеек первой случайной величины;
массив2 – второй интервал ячеек со значениями второй случайной величины.

Пример – Имеются результаты семимесячных наблюдений реализации путевок двух туристских маршрутов тура А и тура В (таблица 4.1).

Таблица 4.1 – 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 поле 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*, т. е. ввести ссылку 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, т. е. 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 4.1);

– нажать кнопку *ОК*.

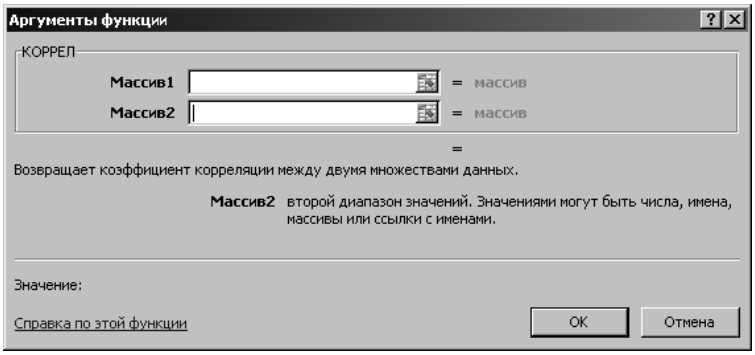


Рисунок 4.1 – Пример установки параметров корреляционного анализа

Практические задания

1 Определить, влияет ли фактор образования на уровень зарплаты сотрудников в гостинице на основании данных таблицы 4.2.

2 Определить, имеется ли взаимосвязь между рождаемостью и смертностью (количество на 1000 чел.) в Санкт-Петербурге (таблица 4.3).

3 Определить, имеется ли взаимосвязь между годовым уровнем инфляции (%), ставкой рефинансирования (%) и курсом доллара (р./долл.), по данным ежегодных наблюдений (таблица 4.4).

Таблица 4.2 – Данные по зарплате сотрудников

Номер варианта	Данные						
	Образование	Зарплата сотрудника					
1	2	3					
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

Окончание таблицы 4.2

1	2	3					
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

Таблица 4.3 – Данные по рождаемости и смертности

Номер варианта	Данные		
	Год	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

Продолжение таблицы 4.3

Номер варианта	Данные		
	Год	Рождаемость	Смертность
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

Окончание таблицы 4.3

Номер варианта	Данные		
	Год	Год	Год
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

Таблица 4.4 – Данные ежегодных наблюдений

Номер варианта	Данные		
	Уровень инфляции	Ставка рефинансирования	Курс доллара
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

Окончание таблицы 4.4

Номер варианта	Данные		
	Уровень инфляции	Ставка рефинансирования	Курс доллара
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

Контрольные вопросы

1 Что такое корреляционный анализ и какие цели он преследует при обработке данных?

2 Каким образом можно загрузить данные для корреляционного анализа в программе R?

3 Какие функции используются для расчета коэффициентов корреляции в R? Приведите примеры использования.

4 Как оценить значимость корреляционных коэффициентов при помощи среды программирования R?

5 Какие методы в R используются для визуализации корреляционных связей между переменными?

6 Как рассчитать и интерпретировать коэффициент корреляции Пирсона с помощью R?

7 Как провести анализ статистической значимости корреляционных коэффициентов в R?

8 Каким образом можно учесть возможную влиятельность выбросов при проведении корреляционного анализа в R?

9 Какие могут быть проблемы и ограничения при использовании корреляционного анализа в среде программирования R?

10 Какие шаги следует предпринять для интерпретации результатов корреляционного анализа, полученных средствами R?

5 Лабораторная работа № 5. Проведение регрессионного анализа средствами среды программирования R

Цель работы: изучение основ регрессионного анализа средствами среды программирования R и сравнение с результатами в MS Excel.

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 параметров. 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 R проведение регрессионного анализа осуществляется с использованием различных пакетов, таких как "lm" (Linear Models) или "glm" (Generalized Linear Models). Основные шаги проведения регрессионного анализа средствами программирования R.

1 Загрузка данных. Исходные данные загружаются в среду программиро-

вания R с помощью функций, таких как read.csv() для файлов CSV или read.table() для текстовых файлов.

2 Предварительный анализ данных. Проводится изучение структуры данных, отображение первых строк, а также оценка основных статистик с использованием функций summary(), str() и др.

3 Построение модели. С использованием функции lm() строится модель линейной регрессии. Например, lm(y ~ x1 + x2, data = mydata), где y – зависимая переменная; x1, x2 – независимые переменные; mydata – набор данных.

4 Оценка модели. Проводится анализ значимости коэффициентов регрессии, коэффициента детерминации, а также проверка предпосылок регрессионной модели.

5 Предсказания. С использованием полученной модели можно провести прогнозирование значений зависимой переменной для новых значений независимых переменных.

6 Визуализация результатов. Для визуализации результатов регрессионного анализа используются различные графические методы, такие как графики остатков, графики зависимости и др.

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*, т. е. ввести ссылку на диапазон анализируемых зависимых данных, содержащий один столбец данных;

– указать *Входной интервал X*, т. е. ввести ссылку на диапазон независимых данных, содержащий до 16 столбцов анализируемых данных;

– указать выходной диапазон, т. е. ввести ссылку на ячейки, в которые будут выведены результаты анализа (рисунок 5.1);

– 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 результаты 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.

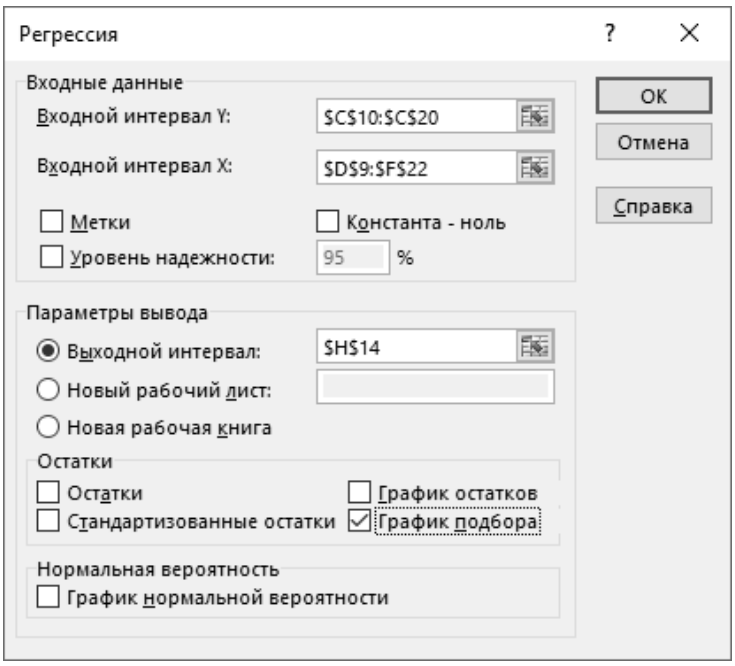


Рисунок 5.1 – Пример заполнения диалогового окна Регрессия

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. Далее (таблица 5.1) приведены стоимость пачки стирального порошка (в рублях) и соответствующий курс доллара (р./долл.).

Таблица 5.1 – 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 урав-

нение, 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 5.2).

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*	*татисти*	*-Значени*	*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

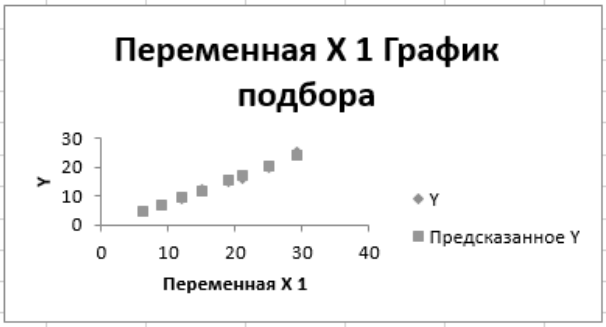


Рисунок 5.2 – Результаты анализа и график соответствия экспериментальных точек и предсказанных по регрессионной модели

Интерпретация результатов. В таблице *Дисперсионный анализ* 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), т. е. $p = 0{,}000000158$ 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).

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. р./долл.).

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 р./долл. ожидаемая стоимость пачки порошка равна 28,8 р.

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 5.2). Фирма совсем молодая, была создана шесть недель назад, и поэтому в Вашем распоряжении имеется статистика только за этот весьма ограниченный промежуток времени.

Таблица 5.2 – Значения наблюдаемых величин

Наблюдаемая величина	Значение					
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:

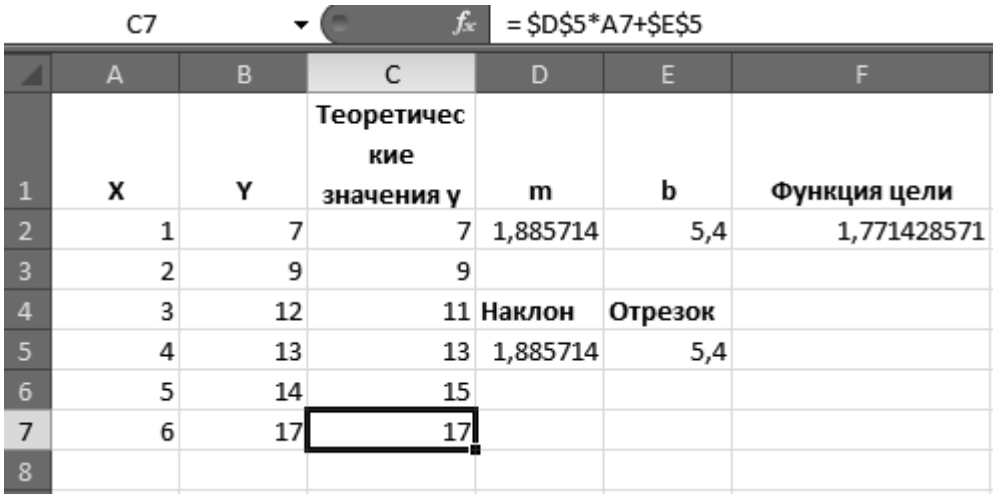
$$z = \sum_{i=1}^{n} (y_i - mx_i - b)^2,$$

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

Для решения этой задачи необходимо выполнить следующие действия.

1 Заполнить ячейки A2:B7 (рисунок 5.3).

2 Отвести под переменные m и b ячейки D2 и E2.



	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 5.3 – Теоретическое значение наблюдаемой величины и коэффициенты уравнения регрессии

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*. 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 (см. рисунок 5.3).

Практические задания

1 Построить зависимость зарплаты (р.) от возраста сотрудника гостиницы по данным таблицы 5.3.

Таблица 5.3 – Данные о зарплате и возрасте сотрудников гостиницы

Номер варианта	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 (таблица 5.4).

Таблица 5.4 – Данные о емкости легких, росте и возрасте для группы мужчин

Номер варианта	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

Продолжение таблицы 5.4

Номер варианта	Данные		
	X_1	X_1	X_1
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

Окончание таблицы 5.4

Номер варианта	Данные		
	X_1	X_1	X_1
10	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 по данным таблицы 5.4.

4 Имеются данные о цене на нефть *x* (д. е.) и индексе акций нефтяных компаний *y* (у. е.), представленные в таблице 5.5.

Таблица 5.5 – Данные о цене на нефть и индексе акций нефтяных компаний

Номер варианта	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.

Контрольные вопросы

1 Что такое регрессионный анализ и для чего он используется?

2 Каким образом можно выполнить регрессионный анализ с помощью R?

3 Какие пакеты в R используются для проведения регрессионного анализа?

4 Каким образом можно изучить связь между двумя переменными при помощи регрессионного анализа?

5 Каким способом можно оценить коэффициенты регрессии в R?

6 Как определить значимость регрессионной модели и ее параметров в R?

7 Что такое множественная регрессия и как ее провести в R?

8 Какой показатель используется для измерения качества регрессионной модели в R?

9 Как провести предсказание на основе регрессионной модели в R?

10 Как можно выполнить визуализацию результатов регрессионного анализа в R?

6 Лабораторная работа № 6. Анализ временных рядов и прогнозирование средствами среды программирования R

Цель работы: изучение моделей временных рядов и методов их анализа средствами среды программирования R.

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.

Анализ временных рядов является важной задачей в статистике и аналитике данных. Он позволяет выявить закономерности и тенденции в изменении переменных во времени и использовать их для прогнозирования будущих значений. В программной среде R существует множество инструментов и пакетов, специально разработанных для анализа временных рядов и прогнозирования.

Основные этапы анализа временных рядов в R включают в себя:

1) загрузка данных. Импорт временных рядов из различных источников, таких как CSV-файлы, базы данных и т. д.;

2) предварительный анализ. Просмотр и визуализация временных рядов, выявление трендов, сезонности, цикличности и выбросов;

3) моделирование. Построение статистических моделей, таких как ARIMA (авторегрессионная интегрированная скользящая средняя), SARIMA (сезонная авторегрессионная интегрированная скользящая средняя), экспоненциальное сглаживание и др.;

4) оценка моделей. Проверка моделей на адекватность, статистические критерии и выбор лучшей модели;

5) прогнозирование. Построение прогнозов будущих значений временного ряда с помощью выбранной модели;

В R для анализа временных рядов и прогнозирования используются

специальные пакеты, такие как "forecast", "TSA", "tseries", "xts", "zoo" и др. Они предоставляют широкий спектр функций для работы с временными рядами, включая методы для визуализации, моделирования и прогнозирования.

Анализ временных рядов и прогнозирование средствами R позволяют выявить скрытые закономерности и использовать их для более эффективного управления и принятия решений в различных областях, таких как финансы, экономика, производство, здравоохранение и др.

Практическое задание

Положение на рынке местного производителя продукта П1 представляют значения продаж по кварталам в таблице 6.1.

Таблица 6.1 – Исходные данные для решения практического задания

Продажа 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 Вычислить сезонные индексы для данных из таблицы 6.1 (использовать центрированные средние значения за четыре квартала).

2 Исключить сезонную составляющую из этих данных.

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

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

Контрольные вопросы

1 Что такое временной ряд и какие основные характеристики он обычно имеет?

2 Какие методы анализа временных рядов используются в программировании R?

3 Каковы основные шаги анализа временных рядов в R?

4 Какие графические методы анализа временных рядов можно применить средствами R?

5 Какие функции R используются для преобразования временных рядов перед прогнозированием?

6 Как проводится прогнозирование временных рядов средствами R?

7 Как провести оценку точности прогнозирования временных рядов в R?

8 Как учесть сезонность и тренд при анализе временных рядов средствами R?

9 Какие пакеты R наиболее эффективны для анализа временных рядов и прогнозирования?

10 Какие специальные инструменты и библиотеки R помогают визуализировать результаты анализа временных рядов?

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 : 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. Л. И. Ниворожкиной. – М. : РИОР; ИНФРА-М, 2016. – 333 с.

4 **Маталыцкий, М. А.** Теория вероятностей, математическая статистика и случайные процессы : учеб. пособие / М. А. Маталыцкий, Г. А. Хацкевич. – Минск : Выш. шк., 2012. – 720 с.

5 **Карманов, Ф. И.** Статистические методы обработки экспериментальных данных с использованием пакета MathCad: учеб. пособие / Ф. И. Карманов. – М. : КУРС; ИНФРА-М, 2019. – 208 с.

6 Практикум по статистике / Под ред. Е. Г. Борисовой. – 4-е изд., перераб. и доп. – М. : МГИМО-Университет, 2020. – 166 с.