

**А.П. КУЛАИЧЕВ**

## ***МЕТОДЫ И СРЕДСТВА КОМПЛЕКСНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ***

**2016 г.**

***Практикум – учебные примеры***  
(файлы данных расположены в папке DAT)

### ***3.5. Пропуски и выбросы***

---

#### ***Пример 1***

***З а д а ч а.*** В эксперименте были измерены значения четырех показателей у восьми животных, при этом некоторые значения не были зафиксированы (табл. 3.1, файл MIS):

***Таблица 3.1. Измерения четырех показателей у 8 животных с пропусками***

Объект	x1	x2	x3	x4
1	1.1	44.3	-2	0.3
2	?	48.1	?	0.97
3	2.8	51	0.1	0.78
4	8.399	?	44	-0.56
5	?	35	?	2
6	5.5	33	2	?
7	5	32.3	1.7	-3.4
8	8.899	-1.1	0	1.1

Произведем анализ пропущенных значений и выбросов:

***Результат:***

Пропущен=6,      Выбросов=3

2	1	2	3
4	-?-		-?-
5		-?-	2
	-?-		-?-

6		-?-
7		-2.1
8	-2.1	

**Выводы:** Выявлено наличие выбросов: в 8-м значении переменной  $x_2$ , в 4-м значении переменной  $x_3$  и в 7-м значении переменной  $x_4$  с амплитудой  $-2,1$ ,  $2$  и  $-2,1$  стандартных отклонений.

Произведем замену пропущенных значений по методу средних и по методу регрессии, совместив их для наглядности (результаты выделены жирным шрифтом).

### Результат:

	Метод средних				Метод регрессии			
	$x_1$	$x_2$	$x_3$	$x_4$	$x_1$	$x_2$	$x_3$	$x_4$
1	1.1	44.3	-2	0.3	1.1	44.3	-2	0.3
2	<b>5.28</b>	48.1	<b>7.63</b>	0.97	<b>2.46</b>	48.1	<b>7.633</b>	0.97
3	2.8	51	0.1	0.78	2.8	51	0.1	0.78
4	8.39	<b>34.65</b>	44	-0.56	8.399	<b>8.504</b>	44	-0.56
5	<b>5.28</b>	35	<b>7.63</b>	2	<b>4.238</b>	35	<b>7.633</b>	2
6	5.5	33	2	<b>0.17</b>	5.5	33	2	<b>0.289</b>
7	5	32.3	1.7	-3.4	5	32.3	1.7	-3.4
8	8.89	-1.1	0	1.1	8.899	-1.1	0	1.1

## Пример 2

**Задача.** На кафедре ВНД МГУ исследовался условный рефлекс на болевой стимул у кошки с измерением его латентного периода до и после введения блокатора  $DI$ -рецепторов SCH 23390 (табл. 3.2, файл LATPER)<sup>1</sup>:

Таблица 3.2. Латентность реакции кошки на стимул до и после введения блокатора боли

До	575	650	450	475	550	500	500	500	500	650	525	550	550	500
	625	2525	600	475	500	450	550	1850	500	475	575			
После	550	625	650	750	575	700	600	775	775	750	625	1050	600	900
	700	525	625	1125	700	725	825							

Попробуем выявить различия между средними значениями латентных периодов по параметрическому критерию Стьюдента (см. разд. 6.4) и непараметрическому критерию Вилкоксона (см. разд. 7.2).

### Результаты:

КРИТЕРИЙ ФИШЕРА И СТЬЮДЕНТА. Файл:latper Переменные: до, после  
Статистика Стьюдента=0.574, Значимость=0.5757, степ.своб = 44  
Гипотеза 0: <Нет различий между выброчными средними>

КРИТЕРИЙ СДВИГА (ПОЛОЖЕНИЯ). Файл:latper Переменные: до, после  
Вилкоксон=406,  $Z=4.016$ , Значимость=2.987E-5, степ.своб = 25,21

Гипотеза 1: <Есть различия между медианами выборок>  
Ван дер Варден=-11.3,  $Z=-3.58$ , Значимость=0.00017, степ.своб=25,21  
Гипотеза 1: <Есть различия между медианами выборок>

<sup>1</sup> Данные предоставлены В.И. Майоровым.

**В ы в о д ы:** Анализ дает нам парадоксальные результаты. Критерий Стьюдента не обнаруживает различий с высокой степенью достоверности ( $P=0,5757$ ), а два непараметрических критерия выявляют различия с высоким уровнем значимости, близким к нулю. В чем здесь дело?

Если же мы повнимательнее присмотримся к исходным данным, то обнаружим в них несколько значений, заметно отличающихся от большинства других по величине. Для более точного исследования этих отклонений выполним процедуру поиска пропусков и выбросов:

**Результаты:**

Пропущн=0,	Выбросов=4
i	до после
12	2.1
16	3.9
18	2.6
22	2.5

**В ы в о д ы:** В исходных данных присутствуют 4 выброса, величиной более двух стандартных отклонений, что привело к сильному сдвигу средних значений двух выборок, поэтому чувствительный к выбросам критерий Стьюдента не обнаружил их различий. Следовательно, необходимо удалить эти экстремальные значения из исходных данных и повторить анализ.

**Результаты:**

КРИТЕРИЙ ФИШЕРА И СТЬЮДЕНТА. Файл: latper

Переменные: до, после

Статистика Стьюдента=5.872, Значимость=1.523E-5, степ.своб = 40

Гипотеза 1: <Есть различия между выборочными средними>

**В ы в о д ы:** После удаления выбросов критерий Стьюдента не подтверждает нулевую гипотезу об отсутствии различий средних, так как ранее это показали нечувствительные к выбросам непараметрические критерии.

## 6.1. Описательная статистика

### Пример

**З а д а ч а.** В антропометрическом исследовании было случайным образом отобрано 50 мужчин и 50 женщин в возрасте от 15 до 70 лет, у которых были зарегистрированы следующие показатели: рост, вес, возраст, цвет глаз и цвет волос (файл PEOPLE). Характер распределения показателей роста, веса и возраста в выборке иллюстрирует рис. 6.1, а соотношения *вес–рост* и *вес–возраст* показаны на диаграммах рассеяния (рис. 6.2).

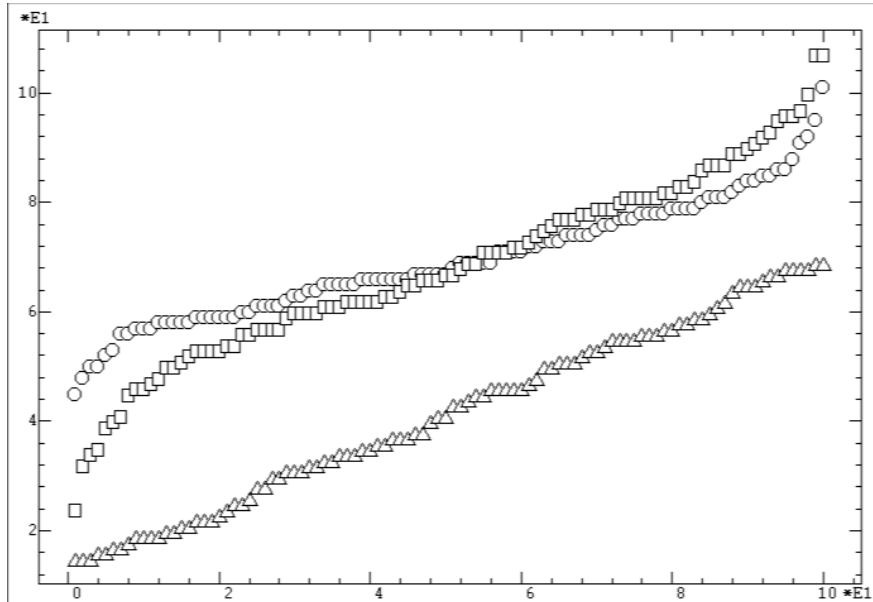


Рис. 6.1. График Кетле для распределения значений антропометрических показателей: квадраты — *рост-100*; круги — *вес*; треугольники — *возраст*; по горизонтальной оси — номер измерения в упорядоченном по возрастанию значений ряду

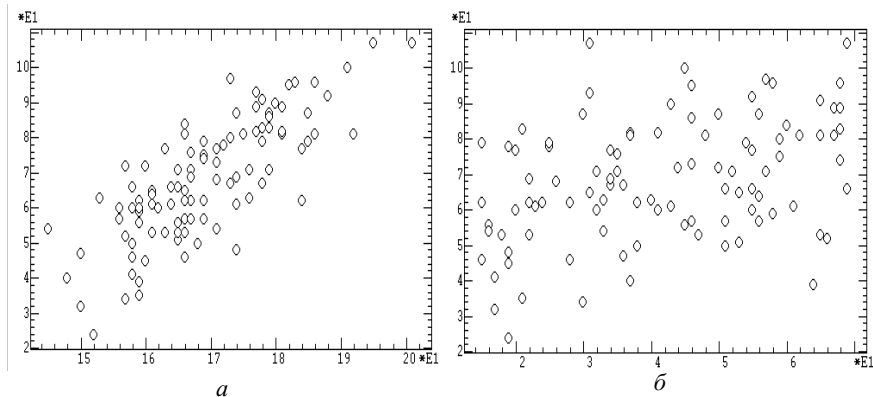


Рис. 6.2. Диаграммы рассеяния антропометрических показателей; разметка осей произведена в десятках единиц измерения:  
 $a$  — *вес* ( $Y$ ) — *рост* ( $X$ );  $b$  — *вес* ( $Y$ ) — *возраст* ( $X$ )

На предварительном этапе исследования необходимо вычислить оценки описательной статистики для этих показателей.

### Результаты:

ОПИСАТЕЛЬНАЯ СТАТИСТИКА.

Файл: people.std



Перемен.	Размер	<-Диапазон->		Среднее-Ошибка		Дисперс	Ст.откл	Сумма
рост	100	145	201	169	1.09	119	10.9	1.69E4
вес	100	24	107	68.3	1.71	293	17.1	6.83E3
возраст	100	15	69	41.7	1.65	272	16.5	4.17E3
Переменная	Медиана	<-Квартили->		ДовИнтСр	<-ДовИнтДисп->		Ош.	СтОткл
рост	168	161	178	2.13	89.6	161	2.9	
вес	67	57	81	3.35	221	398	4.56	
возраст	42	28	55.8	3.23	205	369	4.39	
Переменная	Асимметр.	Значим	Эксцесс	Значим				
рост	0.323	0.087	2.85	0.42				
вес	-0,054	0.41	2.67	0.273				
возраст	0.0072	0.488	1.79	0.0057				

**В ы в о д ы:** В данной выборке наблюдаются следующие средние показатели для мужчин и женщин: вес = 68.3 кг, рост = 169 см, возраст = 41,7 года, при стандартных отклонениях в 10,9 кг, 17,1 см, 16.5 лет. По оценкам асимметрии и эксцесса нормальному распределению соответствует показатели роста и веса (оба уровня значимости существенно больше 0,05), но по оценке асимметрии проходит и показатель возраста. Можно также отметить заметную асимметрию роста, что видимо определено систематически меньшим ростом у женщин (т. е. по этому показателю в выборке присутствуют две различающиеся популяции). С другой стороны, вес не имеет заметной асимметрии, что можно объяснить большим средним весом у женщин, несмотря на их меньший рост.

Продолжение анализа данного примера см. в следующем разделе.

## 6.2. Гистограмма и проверка распределения на нормальность

### Пример 1

**З а д а ч а.** В антропометрическом исследовании ста человек фиксировались, в частности, их вес и возраст (файл PEOPLE). В примере к разд. 6.1 было показано, что по оценкам асимметрии и эксцесса нормальному распределению соответствуют показатели роста и веса. Возраст же не соответствует нормальному распределению по показателю эксцесса. Для более наглядного представления о характере распределений этих показателей полезно построить гистограммы и провести более тщательную проверку соответствия их нормальному распределению. Сделаем это для показателей веса и возраста.

### Результаты:

ГИСТОГРАММА И ТЕСТ НОРМАЛЬНОСТИ. Файл: people.std					
			Переменная: вес		
X-лев.	X-станд	Частота	%	Накопл.	%
24	-2.59	2	2	2	2

33.2	-2.05	5	5	7	7
42.4	-1.51	8	8	15	15
51.7	-0.97	18	18	33	33
60.9	-0.432	21	21	54	54
70.1	0.107	18	18	72	72
79.3	0.646	15	15	87	87
88.6	1.18	10	10	97	97
97.8	1.72	3	3	100	100
107	2.26				

Колмогоров=0,0547, Значимость=0.977, степ.своб = 100

Гипотеза 0: <Распределение не отличается от нормального>

Омега-квадрат=0.0379, Значимость=1.04, степ.своб = 100

Гипотеза 0: <Распределение не отличается от нормального>

Хи-квадрат=1.94, Значимость=0.926, степ.своб = 6

Гипотеза 0: <Распределение не отличается от нормального>

ГИСТОГРАММА И ТЕСТ НОРМАЛЬНОСТИ. Файл: people.std

Переменная: возраст

Х-лев.	Х-станд	Частота	%	Накопл.	%
15	-1.62	16	16	16	16
21	-1.25	8	8	24	24
27	-0.89	11	11	35	35
33	-0.525	12	12	47	47
39	-0.161	8	8	55	55
45	0.203	12	12	67	67
51	0.567	13	13	80	80
57	0.931	7	7	87	87
63	1.29	13	13	100	100
69	1.66				

Колмогоров=0.0809, Значимость=0.128, степ.своб = 100

Гипотеза 0: <Распределение не отличается от нормального>

Омега-квадрат=0.172, Значимость=0.0115, степ.своб = 100

Гипотеза 1: <Распределение отличается от нормального>

Хи-квадрат=41.7, Значимость=6.72E-6, степ.своб = 6

Гипотеза 1: <Распределение отличается от нормального>

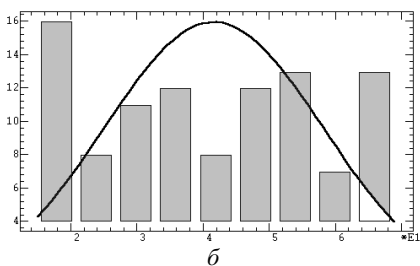
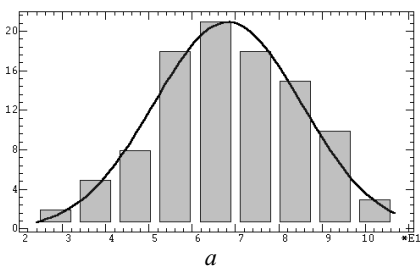


Рис. 6.7. Гистограммы выборочных распределений  
с графиками плотности вероятности нормального распределения:  
а — вес в выборке из 100 человек; б — возраст в той же выборке

**В ы ы о д ы:** Согласно результирующим уровням значимости всех трех критериев ( $P \gg 0,05$ ) можно принять гипотезу о нормальном распределении веса людей (рис. 6.7, а).

Что же касается возраста, то по двум критериям (чувствительным к различиям «на концах») его распределение не соответствует нормальному закону (на глаз это хорошо видно по рис. 6.7, б). Это соответствует содержательным соображениям о том, что в случайной выборке возраст

людей должен быть распределен достаточно равномерно. Это предположение будет проверено при рассмотрении примера 3 разд. 7.1.

Продолжение анализа данного примера см. в следующем разделе.

## Пример 2

**З а д а ч а.** Очень полезной практикой для выработки внутреннего «ощущения случайности» случайных процессов, а также для понимания сравнительной чувствительности вышерассмотренных критериев является следующее простое упражнение. Сгенерируйте несколько случайных выборок согласно нормальному закону распределения достаточно большого размера (операция «Генератор чисел» из Блока преобразований, разд. 3.4), проверьте их распределение на нормальность и визуально сравните их гистограммы.

Ниже приведены результаты проверки по критерию Колмогорова четырех подобных выборок, включающих по 100 элементов.

### Результаты:

Колмогоров=0.073, Значимость=0.259, степ.своб = 100  
 Колмогоров=0.049, Значимость=0,058, степ.своб = 100  
 Колмогоров=0.048, Значимость=0.242, степ.своб = 100  
 Колмогоров=0,057, Значимость=0.076, степ.своб = 100

**В ы ы о д ы:** Как видно из полученных результатов, все выборки подчиняются нормальному закону распределения, но уровни значимости нулевой гипотезы колеблются в довольно широких пределах, иногда приближаясь вплотную к критической границе.

В качестве самостоятельного упражнения можно сравнить результаты по другим критериям, а также их зависимость от размеров выборок.

## 6.3. Линейная корреляция

### Пример 1

**З а д а ч а.** Целью исследования было выявление зависимости между продуктивностью пшеницы и картофеля на соседних полях. Для этого были использованы данные по урожайности [ц/га] за 12 последовательных лет (табл. 6.3.1, переменные LC1, LC2 в файле TST)

Таблица 6.3.1. Урожайность пшеницы и картофеля за 12 лет [ц/га]

Пшен	20,1	23,6	26,3	19,9	16,7	23,2	31,4	33,5	28,2	35,3	29,3	30,5
Карт	7,2	7,1	7,4	6,1	6	7,4	9,4	9,2	8,8	10,4	8	9,7

Для выявления связи этих двух показателей данные были подвергнуты корреляционному анализу.

### Результаты:

ПАРАМЕТРИЧЕСКАЯ КОРРЕЛЯЦИЯ. Файл: tst.std Переменные: lc1, lc2  
 Коэфф.корреляции=0.8516 Т=5.138, Значимость=0.0006, степ.своб = 10  
 Гипотеза 0: <Коэффициент корреляции отличен от нуля>

**В ы в о д ы:** Проверка нулевой гипотезы (уровень значимости равен 0,0006, что существенно меньше 0,05) выявляет значимую корреляцию между урожайностью пшеницы и картофеля.

## Пример 2

**З а д а ч а.** Целью исследования было выявление связей между антропометрическими и умственными показателями детей. В экспериментах у 22 детей (табл. 6.3.2, файл COR1) измерялись: время решения математических задач, возраст и размер стопы.

Таблица 6.3.2. Результаты исследования детей

Ученик	Время решения задачи, мин	Возраст, лет	Размер стопы, см
1	4,6	10,1	17,2
2	2,8	10,6	18
3	5,4	10,3	17
4	4	10,1	16,7
5	2,6	10,9	17,8
6	4,4	10,2	17,4
7	4,6	9,9	16,5
8	4,4	10,1	16,5
9	5,2	9,3	15,1
10	2,8	10,8	17,8
11	5,2	9,4	15,5
12	3,6	10,2	16,9
13	3,8	9,8	16,9
14	6	9,1	15
15	4,2	10,1	16,7
16	4,4	10,2	17
17	4	10,2	16,5
18	4,4	10,1	17,2
19	4,2	10	16,1
20	3,6	10,3	16,7
21	3,4	10,5	17,2
22	2	11,1	20

Для выявления связи этих признаков данные были подвергнуты корреляционному анализу.

### Результаты:

ПАРАМЕТРИЧЕСКАЯ КОРРЕЛЯЦИЯ. Файл: cor1.std  
 Корреляционная матрица  
 время возраст  
 возраст -0.869  
 стопа -0.822 0.906  
 Критическое значение=0.419  
 Число значимых коэффициентов=3 (100%)  
 Частные корреляции  
 время возраст

---

```

возраст    -0.515
стопа      -0.163    0.682
Критическое значение=0.429
Число значимых коэффициентов=2 (66%)

```

**В ы в о д ы:** Расчет коэффициентов парной корреляции выявил удивительный факт — наличие высокой обратной связи между умственными способностями и размером стопы ребенка (корреляция  $-0,8538$  по абсолютной величине значительно превышает критическое значение  $0,4186$ ). Однако расчет частных корреляций показал, что эти признаки не связаны друг с другом (частная корреляция  $-0,1637$  по абсолютной величине много ниже критического уровня  $0,429$ ), поскольку оба они сильно зависят от возраста.

---

### Пример 3

**Задача.** Продолжим анализ антропометрических данных из примеров к разд. 6.1, 6.2 (файл PEOPLE) с целью исследования корреляций между показателями веса, роста и возраста людей. Уже визуальный анализ диаграмм рассеяния (см. рис. 6.2) выявляет хорошую коррелированность веса и роста и меньшую, но все же заметную корреляцию между весом и возрастом.

**Результаты** (сокращенно):

```

ПАРАМЕТРИЧЕСКАЯ КОРРЕЛЯЦИЯ.  Файл: cor1.std
Корреляционная матрица
      рост      вес
вес    0.784
возраст 0.106    0.385
Критическое значение=0.194
Число значимых коэффициентов=2 (66%)

```

**В ы в о д ы:** Значимые корреляции обнаружены между весом и ростом, а также между весом и возрастом, хотя последняя связь оказалась в 2 раза слабее первой (как мы и предполагали выше). Связь между возрастом и ростом мала и незначима. Она, видимо, является следствием продолжения роста человека от 15 до 25 лет, что компенсируется небольшим уменьшением роста в преклонном возрасте. Анализ частных коэффициентов корреляции здесь не проводится, поскольку нет причин считать наличие сильного влияния связей с третьим показателем на взаимную связь двух других.

Продолжение анализа данного примера см. в следующем разделе.

## 6.4. Критерии Стьюдента и Фишера

---

### Пример 1

**Задача.** Целью исследования было сравнение влияния на урожайность пшеницы двух агротехнических методов, применяемых на двух соседних полях в течение 10 лет (табл. 6.4.1, переменные  $f1, f2$  в файле TST).

Таблица 6.4.1. Урожайность пшеницы за 10 лет [ц/га] на соседних полях при двух агротехнических методах

метод1	20	17,9	20,6	22	21,4	23,8	21,4	19,8	18,4	22,5
метод2	22,1	18,5	19,4	22,1	21,7	24,9	21,6	20,3	18,7	23,1

### Результаты:

КРИТЕРИИ ФИШЕРА И СТЬЮДЕНТА. Файл: tst.std Переменные: f1, f2  
Статистика Фишера=0.81, Значимость=0.379, степ.своб = 9,9

Гипотеза 0: <Нет различий между выборочными дисперсиями>  
Статистика Стьюдента=0.534, Значимость=0.606, степ.своб = 18  
Разность средних=0.46, доверит.интервал=0.522

Гипотеза 0: <Нет различий между выборочными средними>  
Стьюдент для парных данных=1.76, Значимость=0.11, степ.своб = 9  
Гипотеза 0: <Нет различий между выборочными средними>

**Выводы:** Как можно видеть из полученных результатов анализа, ни критерий Стьюдента, ни критерий Фишера не выявляют заметных различий между рассматриваемыми методами сбора урожая (полученные уровни значимости 0,379 и 0,6056 существенно больше 0,05).

Отметим также, что в данном случае выборки можно рассматривать как парные переменные, поскольку поля являются соседними и не отличаются ни по почве, ни по климату, ни по другим условиям, а измерения проводились синхронно по годам. Однако даже более чувствительный к различиям парный критерий Стьюдента не выявляет таковых.

Количественную оценку диапазона значений для генеральной разности средних относительно выборочной разности 0.46 дает доверительный интервал 0,522, который включает и нулевое значение с перекрытием в 0,522–0,46=0,062, т. е. с «запасом» на  $0,062/0,46 \cdot 100 = 13,5\%$ .

---

### Пример 2

**Задача.** Продолжим исследование антропометрических данных из примеров к разд. 6.1–6.3 (файл PEOPLE) с целью исследования различий между показателями веса, роста и возраста мужчин и женщин. Перед этим

необходимо произвести ряд преобразований, чтобы разделить данные по мужчинам и женщинам (см. пример к разд. 3.4)

### **Результаты** (сокращенно):

КРИТЕРИЙ ФИШЕРА И СТЬЮДЕНТА. Файл: people.std  
 Переменные: рост-м, рост-ж  
 Статистика Стьюдента=3.97, Значимость=0.000327, степ.своб = 98  
 Гипотеза 1: <Есть различия между выборочными средними>  
 Разность средних=8.03, доверит.интервал=0.000661  
 С поправкой Бонферрони: значимость=0.000259, степ.своб=294,  
 крит.значимость=0.00333  
 Гипотеза 1: <Есть различия между выборочными средними>  
 Переменные: вес-м, вес-ж  
 Статистика Стьюдента=4.33, Значимость=0.000141, степ.своб = 98  
 Гипотеза 1: <Есть различия между выборочными средними>  
 Разность средних=13.7, доверит.интервал=0.000444  
 С поправкой Бонферрони: значимость=0.000107, степ.своб=294,  
 крит.значимость=0.00333  
 Гипотеза 1: <Есть различия между выборочными средними>  
 Переменные: возраст-м, возраст-ж  
 Статистика Стьюдента=0.141, Значимость=0.883, степ.своб = 98  
 Гипотеза 0: <Нет различий между выборочными средними>  
 Разность средних=0.468, доверит.интервал=2.93

**Выводы:** Результаты показывают наличие различий в весе и росте по признаку пола (уровни значимости намного меньше критического уровня, да и доверительные интервалы составляют незначительный процент от разностей средних), но не в возрасте (уровень значимости нулевой гипотезы очень высок, а доверительные интервалы в шесть раз превышают разность средних). В данном случае поправку Бонферрони учитывать не следует, поскольку здесь имеет место не сравнение различий нескольких групп по одному показателю, а различие двух групп по трем разным показателям.

Продолжение анализа данного примера см. разд. 7.6.

---



---

## **НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ**

### **7.1. Критерий хи-квадрат**

---

#### **Пример 1**

**Задача.** Необходимо сравнить частоты событий, измеренные в эксперименте, с теоретическими частотами (табл. 7.1.1, переменные Chi1 и Chi3 в файле NPT).

*Таблица 7.1.1. Экспериментальные и теоретические частоты событий*

---

Эксперимент	7	11	13	19	16	7	7
Теория	19	16	7	11	13	7	7

### Результаты:

КРИТЕРИЙ ХИ-КВАДРАТ. Файл: npt.txt

Переменные: chil, chi3

Хи-квадрат=10.71, Значимость=0.0978, степ.своб = 6

Гипотеза 0: <Нет различий между двумя распределениями>

**В ы в о д ы:** Вычисленные уровни значимости критерия хи-квадрат согласия позволяют принять гипотезу о соответствии экспериментальных данных предполагаемому теоретическому распределению ( $P=0,0978$  больше 0,05).

---

## Пример 2

**З а д а ч а.** На предприятии имеются данные о числе работников с заработной платой в заданных пределах для двух возрастных категорий (табл. 7.1.2, переменные CH1, CH2 из файла NPT). Необходимо проверить гипотезу об отсутствии различий в оплате труда между двумя возрастными категориями работников.

Поскольку данные представляют собой гистограммы, то для сравнения используем критерий однородности этих двух экспериментальных распределений, как принадлежащих одной генеральной совокупности.

Таблица 7.1.2. Экспериментальные и теоретические частоты событий

Зарплата	100–120	120–140	140–160	160–180	180–200	200–220
Категория1	71	430	1072	1609	1178	158
Категория2	54	324	894	1202	903	112

### Результаты:

КРИТЕРИЙ ХИ-КВАДРАТ. Файл: npt.txt Переменные: chil, chit

Хи-квадрат=3.218, Значимость=0.0781, степ.своб = 6

Гипотеза 0: <Нет различий между двумя распределениями>

**В ы в о д ы:** Полученный уровень значимости хи-квадрат критерия однородности ( $P=0,781$  существенно больше 0,05) позволяет принять нулевую гипотезу об отсутствии различий в оплате труда.

---

## Пример 3

**З а д а ч а.** Продолжим анализ антропометрических измерений из раздела 6.1. В разд. 6.2 было показано, что показатель возраста в анализируемой выборке имеет распределение, отличающееся от нормального, и было высказано предположение о равномерном его распределении в



представленном диапазоне 15–70 лет. Проверим это предположение. Для этого необходимо при построении гистограммы подтвердить необходимость сохранения ее в матрице данных. Кроме того, принимая во внимание, что объем нашей выборки составляет 100 человек, установим при построении гистограммы число бинов равным 10. Тогда проще будет сформировать вторую переменную, необходимую для данного метода, содержащую гистограмму теоретического распределения с 10 бинами, по 10 элементов в каждом.

### **Результаты:**

КРИТЕРИЙ ХИ-КВАДРАТ. Файл: people.std  
 Хи-квадрат=8.6, Значимость=0.475, степ.своб = 9  
 Гипотеза 0: <Нет различий между двумя распределениями>

**Выводы:** Полученный уровень значимости хи-квадрат критерия согласия позволяет принять нулевую гипотезу об отсутствии различий между распределением выборочных значений показателя возраста и равномерным распределением.

## **7.2. Критерии различия сдвига (положения)**

### **Пример 1**

**Задача.** В эксперименте измерены уровни радиоактивности для двух групп препаратов (табл. 7.2.1, переменные WT1, WT2 из файла NPT).

Таблица 7.2.1. Уровни радиоактивности двух препаратов [имп/с]

Препарат 1	340	343	322	349	332	320	313	304	329
Препарат 2	318	321	318	301	312				

Необходимо оценить достоверность различий между этими препаратами в средних значениях.

### **Результаты:**

КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ). Файл: npt.txt Переменные: wt1, wt2  
 Вилкоксона=82, Z=-1.935, Значимость=0.0264, степ.своб = 9,5  
 Гипотеза 1: <Есть различия между медианами выборок>  
 Ван дер Вардена=2.944, Z=1.902, Значимость=0.0285, степ.своб = 9,5  
 Гипотеза 1: <Есть различия между медианами выборок>

**Выводы:** Применение критериев различия сдвига Вилкоксона и Ван дер Вардена, как можно видеть (полученные уровни значимости 0,0266 и 0,0285 меньше 0,05), позволяют принять гипотезу о различии между препаратами.

## Пример 2

**Задача.** В эксперименте оценивалась светочувствительность куриных эмбрионов в темноте и на свету (по числу клевков по скорлупе). Измерения проводились попеременно у 25 эмбрионов (табл. 7.2.2, переменные SIG1, SIG2 в файле NPT).

Ожидается, что реакции на световой стимул будет соответствовать положительный сдвиг. Необходимо проверить это предположение.

Таблица 7.2.2. Светочувствительность 25 куриных эмбрионов в темноте и на свету  
[числу клевков в минуту по скорлупе]

Свет	6	14	26	7	8	23	11	9	19	26	18	18	18	14
Темнота	5	21	73	25	3	77	59	13	36	46	9	25	59	38

Свет	55	15	30	21	27	8	24	21	18	23	31
Темнота	70	36	55	46	25	30	39	46	71	31	33

### Результаты:

КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ). Файл: npt.txt Переменные: sig1, sig2  
Вилкоксон=455,  $Z=3.54$ , Значимость=0.000199, степ.своб = 25,25

Гипотеза 1: <Есть различия между медианами выборок>

Ван-дер-Варден=-10.7,  $Z=-3.19$ , Значимость=0.0008, степ.своб=25,25

Гипотеза 1: <Есть различия между медианами выборок>

Для парных данных:

Вилкоксон=17.5,  $Z=-3.9$ , Значимость=4.8E-5, степ.своб = 2,25

Гипотеза 1: <Есть различия между медианами выборок>

Знаков=4,  $Z=-3.2$ , Значимость=0.0007, степ.своб = 2,25

Гипотеза 1: <Есть различия между медианами выборок>

**Выводы:** Анализируемые данные представляют парные переменные, поскольку измерения проводились у одних и тех же эмбрионов. Результаты применения критериев сдвига для парных выборок выявляют различие между двумя выборками (вычисленные уровни значимости 4,8E-5 и 0,0007 существенно меньше значения 0,05), что говорит о наличии реакции на свет у эмбрионов.

## Пример 3

**Задача.** Применим данный метод для проверки гипотезы о том, что медиана совокупности, представленной выборкой SIG1 (из Примера 2) имеет значение 20. Для этого введем в переменную MSIG 25 константу со значением 20.

### Результаты:

КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ). Файл: npt.txt Переменные: sig1, msig  
Вилкоксон=600,  $Z=0.7778$ , Значимость=0.2183, степ.своб = 25,25

Гипотеза 0: <Нет различий между медианами выборок>

Ван дерВарден=-1.981,  $Z=-0.5916$ , Значимость=0.277, степ.своб=25,25

Гипотеза 0: <Нет различий между медианами выборок>

Для парных данных:

Вилкоксон=137.5,  $Z=-0.6727$ , Значимость=0.2505, степ.своб = 2,25

Гипотеза 0: <Нет различий между медианами выборок>

Знаков=11,  $Z=-0.4$ , Значимость=0.3445, степ.своб = 2,25

Гипотеза 0: <Нет различий между медианами выборок>

**В ы в о д ы:** Результаты анализа позволяют принять нулевую гипотезу о равенстве медианы значению 20.

## 7.3. Критерии различия масштаба (рассеяния)

### Пример

**З а д а ч а.** Был произведен полумикроанализ на железо 20 препаратов железистой сыворотки (применяется от малокровия) с использованием традиционного и нового методов (табл. 7.3.1, переменные AB1, AB2 в файле NPT).

Таблица 7.3.1. Результаты полумикроанализа на железо 20 препаратов железистой сыворотки

Метод1	111	107	100	99	102	106	109	108	104	99
Метод2	107	108	106	98	105	103	110	105	104	100
Метод1	101	96	97	102	107	113	116	113	110	98
Метод2	96	108	103	104	114	114	113	108	106	99

Внедрение нового метода возможно в случае, когда он не приводит к существенному ухудшению точности измерений, выраженному в возрастании разброса результатов измерений. Необходимо проверить эту гипотезу.

### Результаты:

КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ). Файл: npt.std Переменные: ab1, ab2

Для парных данных:

Вилкоксон=86,  $Z=-0.364$ , Значимость=0.358, степ.своб = 2,19

Гипотеза 0: <Нет различий между медианами выборок>

Знаков=8,  $Z=-0.459$ , Значимость=0.323, степ.своб = 2,19

Гипотеза 0: <Нет различий между медианами выборок>

КРИТЕРИИ МАСШТАБА (РАССЕЯНИЯ). Файл: npt.std Переменные: ab1, ab2

Ансари-Вредли=185.5,  $Z=-1.327$ , Значимость=0.0922, степ.своб=20,20

Гипотеза 0: <Нет различий между выборками в масштабах>

Клотц=19.2,  $Z=0.6874$ , Значимость=0.2459, степ.своб = 20,20

Гипотеза 0: <Нет различий между выборками в масштабах>

**В ы в о д ы:** Применение критериев сдвига (результаты здесь не приводятся) позволяет принять нулевую гипотезу отсутствия различий на уровне значимости 0.02. Поэтому применение критериев масштаба

Ансари–Бредли и Клотца к этим данным допустимо. Их результаты показывают отсутствие достоверных различий в разбросе значений сравниваемых выборок (полученные уровни значимости 0,0922 и 0,246 больше 0,05), т. е. точность измерений не уменьшается.

## 7.4. Критерии интегральных различий

**Назначение.** Критерии этого класса предназначены для обнаружения всех возможных отклонений от гипотезы об идентичности двух совокупностей.

**Действия и результаты.** Для анализа нужно выбрать из электронной таблицы две или несколько переменных (см. рис. 2.3).

Вычисляются значение статистики  $D$  Смирнова (часто ошибочно называемой статистикой Колмогорова–Смирнова) и уровень значимости  $P$  нулевой гипотезы об отсутствии интегральных различий между выборками. Если  $P > 0,05$ , нулевая гипотеза может быть принята.

В случае нескольких выбранных переменных подобные вычисления производятся для всех пар переменных

**Ограничения:** размер выборки должен быть не больше  $l$ , где  $l=16000$ , 5000, 1000, 100 при объеме матрицы данных в 64000, 20000, 4000 и 400 чисел.

**Литература:** [6, 33].

---

### Ф о р м у л ы

$$D = \max_i (F_x(z) - F_y(z)) = \max_i \left( \frac{n_x(z)}{N} - \frac{n_y(z)}{M} \right);$$

где:  $z$  — общая шкала значений переменных  $x$ ,  $y$ ;  $n_x(z)$ ,  $n_y(z)$  — число выборочных значений  $x$ ,  $y$ ; меньших  $z$ ;  $N$ ,  $M$  — размер выборок  $x$ ,  $y$ .

Аппроксимация для  $P$  вычисляется по формуле:  $P = 2e^{-\frac{D^2 mn}{m+n}}$ .

---

### П р и м е р

**З а д а ч а.** Был измерен привес свиней при двух различных рационах питания (табл. 7.4.1, переменные KST1, KST1 в файле NPT):

Таблица 7.4.1. Привес свиней при двух рационах питания

Рацион1	11,5, 26, 29,1, 19,7, 2,3, 22,6, 30,9, 10,8, 23,2, 38,8, 21,5
Рацион2	18,4, 15,5, 25,2, 16,9, 24, 13,3, 17,9, 13,2

Требуется оценить достоверность различий этих двух рационов.

### **Результаты:**

КРИТЕРИЙ КОЛМОГОРОВА–СМИРНОВА. Файл:npt.txt Переменные: kst1, kst2  
Смирнов=0.4773, Значимость=0.2425, степ.своб = 11,8

Гипотеза 0: <Нет интегральных различий между выборками>

**Выводы:** Применение критерия Смирнова позволяет принять гипотезу об отсутствии различий между этими двумя рационами (вычисленный уровень значимости 0,2425 существенно больше 0,05).

## **7.5. Ранговая корреляция**

### **Пример**

**Задача.** Была измерена степень усвояемости (*L*–Хантера) для 9 партий консервированного тунца (переменная NCOR1 в файле NPT) и параллельно были получены результаты опроса по 6–балльной шкале, усредненные по 80 потребителям (переменная NCOR2).

Таблица 7.5.1. Оценки усвояемости для 10 партий консервированного тунца

L–Хантера	44.4	45.9	41.9	53.3	44.7	44.1	50.7	45.2	60.1
Опрос	2.6	3.1	2.5	5	3.6	4	5.2	2.8	3.8

Исходно предполагается, что мера Хантера положительно связана с баллами опроса. Необходимо подтвердить или опровергнуть это предположение.

### **Результаты:**

НЕПАРАМЕТРИЧЕСКАЯ КОРРЕЛЯЦИЯ. Файл: npt.txt

Переменные: ncor1, ncor2

Кенделл=0.4444, Z=1.668, Значимость=0.0476, степ.своб = 9

Гипотеза 1: <Есть корреляция между выборками>

Спирмен=0.6, Z=1.691, Значимость=0.0453, степ.своб = 9

Гипотеза 1: <Есть корреляция между выборками>

**Выводы:** Исходное предположение подтверждается результатами анализа, согласно которым нулевая гипотеза об отсутствии связи между двумя признаками может быть отвергнута на уровнях значимости, равной 0,0476, 0,0453.

**Примечание:** Точные таблицы распределения Кенделла дают для этого примера уровень значимости 0,006 вместо асимптотического 0,0476.

## 7.6. Анализ таблиц сопряженности

### Пример 1

**З а д а ч а.** Лечение почечной недостаточности проводят с периодическим подключением искусственной почки через артериовенозный шунт. Из-за завихрений тока крови места соединения шунта с сосудами часто тромбируются. Считается, что прием аспирина снижает риск тромбоза. Для проверки этого были сформированы две группы больных, принимавших аспирин и не принимавших его (контроль). В обеих группах подсчитывалось число случаев тромбоза. Полученные результаты представлены в табл. 7.6.1 (файл TAB).

Таблица 7.6.1. Число случаев наличия и отсутствия тромбоза сосудов при приеме аспирина и без такового

	Тромбоз есть	Тромбоза нет
Контроль	18	7
Прием аспирина	6	13

Необходимо оценить степень связанности этих двух признаков с построением их диаграммы.

#### Результаты:

КРОССТАБУЛЯЦИЯ. Файл: tab.std

Столбцы: Тромбоз Нет

Строки: Контроль Аспирин

```

Наблюдаемые частоты признаков
18      7      25
 6      13     19
-----|
24      20     44
Процентная встречаемость признаков по рядам:
72      28
31.6    68.4
Процентная встречаемость признаков по столбцам:
75      35
25      65
Общая процентная встречаемость признаков:
40.9    15.9 |    56.8%
13.6    29.5 |    43.2%
-----|
54.5%    45.5%
Ожидаемые частоты признаков:
13.6     11.4
10.4     8.64
Остаточные частоты признаков (набл-ожд):
4.36    -4.36
-4.36    4.36

```

Хи-квадрат = 7.19, Значимость = 0.00734, степ.своб = 1

Гипотеза 1: <Есть связь между признаками>

$V$ -коэфф.Краммера =0.404

Лямбда Гудмана и Крускала: симметр, ряд, столб =0.333, 0.316, 0.35

Тау- $b$  Кенделла =0.402 Тау- $c$  Кенделла =0.397

Гамма Гудмана и Кенделла =0.696  $d(x, y)$  Соммера=0.404, 0.4

**В ы в о д ы:** Вычисленный уровень значимости критерия хи-квадрат позволяет принять гипотезу о зависимости тромбоза от приема или неприема аспирина ( $P=0,007$  меньше 0,05). Значения последующих коэффициентов раскрывают различные аспекты выявленной взаимосвязи этих признаков, согласно их выше рассмотренным свойствам. Диаграмма (рис. 7.2) показывает значительную симметричную связь между признаками: большим случаям тромбоза при контроле соответствуют меньше случаев при приеме аспирина. Эту симметрию отражают и большие значения ряда коэффициентов (лямбда, тау- $c$ , гамма и  $d(x, y)$ ).

Данный пример может быть также проанализирован методом попарного сравнения частот событий (разд. 12.4) в клетках таблицы, для чего надо предварительно разделить число событий на число испытуемых каждой категории.

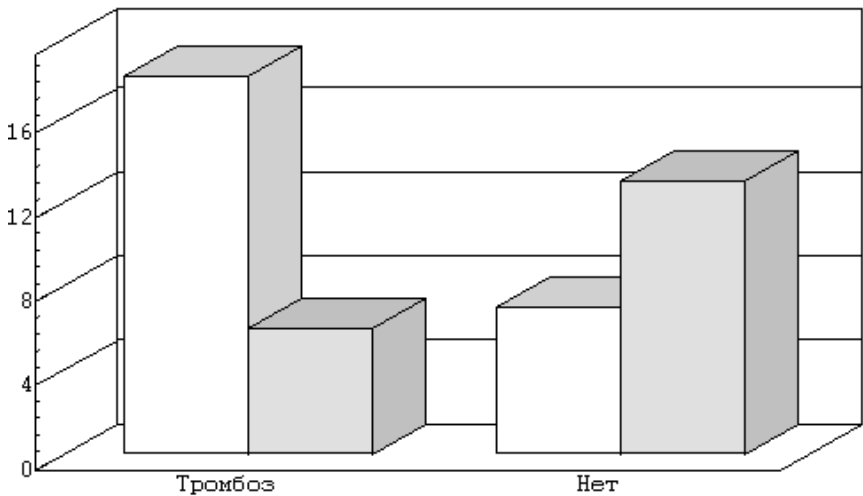


Рис. 7.2. Столбиковая диаграмма кросстабуляции: тромбоз – аспирин

## Пример 2

**Задача.** Продолжим анализ антропометрических данных из примеров к разделам 6.1–6.4 (файл PEOPLE) с целью исследования связи между цветом глаз и цветом волос (предварительно из матрицы данных следует удалить все переменные, кроме двух анализируемых). Здесь мы имеем в качестве исходных данных не готовую таблицу кросстабуляции, а две нативные переменные (глаза, волосы), причем их значения выражены не в

числовой, а в номинальной шкале (синий, зеленый, коричневый, черный, светлый, каштановый).

### Результаты (сокращенно):

КРОССТАБУЛЯЦИЯ. Файл: people.std  
 Столбцы: светлый русский черный коричн  
 Строки: синий серый коричн

Наблюдаемые частоты признаков				
26	10	3	2	41
14	18	11	4	47
2	6	3	1	12
-----				
42	34	17	7	100

Хи-квадрат =14.5, Значимость=0.0241, степ.своб = 6

Гипотеза 1: <Есть связь между признаками>

Коефф. Фи =0.381 Коефф.сопряж. Пирсона =0.356

V-коефф. Граммера =0.27

Ламбда Гудмана и Крускала: симметр,ряд,столб =0.18, 0,0566, 0.293

Тау-b Кенделла =0.312 Тау-c Кенделла =0.297

Гамма Гудмана и Кенделла =0.471 d(x,y)Соммера=0.332, 0.294

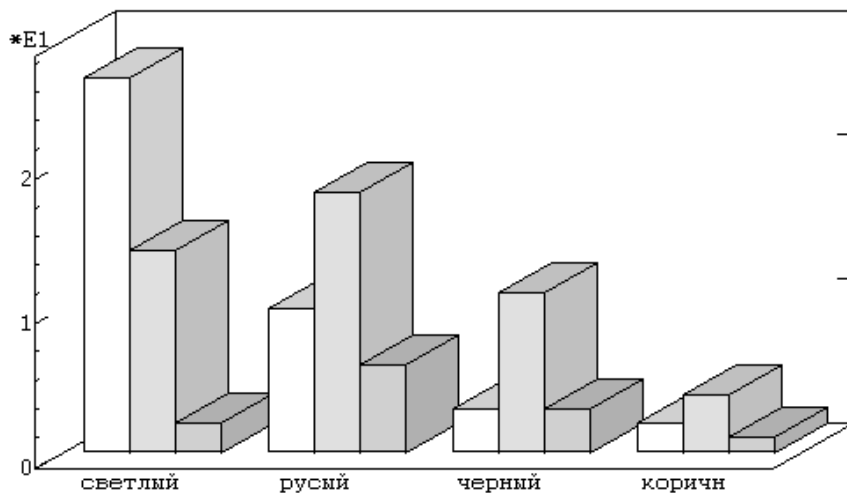


Рис. 7.3. Столбиковая диаграмма кросстабуляции: цвет глаз – цвет волос

**В ы в о д ы:** Результаты анализа свидетельствуют, что имеет место несомненная связь между цветом глаз и цветом волос ( $P=0,024$  меньше 0,05). Диаграмма (рис. 7.3) показывает значительную асимметричность связи между признаками: большие–меньшие встречаемости цвета глаз воспроизводятся для большинства цвета волос. Это отражают и сравнительно меньшие значения коэффициентов симметрии *лямбда*.



---

---

# ДИСПЕРСИОННЫЙ АНАЛИЗ ФАКТОРНЫХ ЭФФЕКТОВ

## 8.2. Однофакторный дисперсионный анализ

### 8.2.1. Параметрические методы

---

#### Пример 1

**Задача.** В исследовании оценивалась урожайность четырех различных сортов пшеницы, выращиваемых на нескольких (различных по числу) участках примерно одного почвенного типа (табл. 8.2.1, файл A1). Необходимо выяснить, отличаются ли эти сорта по урожайности.

Таблица 8.2.1. Урожайность четырех сортов пшеницы [ц/га]

	Сорт 1	Сорт 2	Сорт 3	Сорт 4
Участки	17	15.8	17.4	15.7
	17.2	17	16.6	16.8
	16.1	16.4	16.2	15.1
	17		15.6	15.2
	16.8		15.5	
			17.2	

#### Результаты:

1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: a1.std параметрический

Источник	Сум.квадр	Ст.своб	Ср.квадр	Сила влияния
Факт.1	2.824	3	0.9412	-0.1505
Остат.	6.436	14	0.4597	
Общая	9.26	17	0.5447	

$F(\text{фактор1})=2.047$ ,  $\text{Значимость}=0.1528$ ,  $\text{степ.своб} = 3,14$   
Гипотеза 0: <Нет влияния фактора на отклик>

**Выводы:** Дисперсионный анализ показывает отсутствие заметного влияния на урожайность фактора сорта пшеницы с уровнем значимости 0,152.

## Пример 2

**Задача.** Исследовались показатели прочности шести сплавов (в испытаниях нескольких образцов), из которых четвертый сплав является стандартным (табл. 8.2.2, файл SHEFFE). Следует верифицировать наличие общих различий между всеми сплавами, после чего оценить попарные различия.

Таблица 8.2.2. Показатели прочности шести легированных сплавов

	Сплав1	Сплав2	Сплав3	Стандарт	Сплав4	Сплав5
Образцы	15,1	15,3	15,1	15,3	15,2	15,2
	15	15,5	15,6	14,9	14,8	15,1
	15,4	15,7	15,4	15	15,4	15,3
	15	15,8	15,7	14,9	15	15,1
		15,3		15,2	15	
				14,9		

### Результаты:

1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: sheffe.std  
параметрический

Источник	Сум.кв.др	Ст.своб	Ср.кв.др	Сила влияния
Факт.1	1.013	5	0.2026	0.02548
Остат.	0.9143	22	0.04156	
Общая	1.927	27	0.07138	

F(фактор1)=4.874, Значимость=0.004, степ.своб = 5,22

Гипотеза 1: <Есть влияние фактора на отклик>

Переменные	Разность	Интервал	Значим	Гипотеза H1
Парные сравнения Шеффе				
1-2	0.395	0.4981	0.1836	
1-3	0.325	0.525	0.4326	
1-4	0.09167	0.4793	0.9898	
1-5	0.045	0.4981	0.999	
1-6	0,05	0.525	0.9989	
2-3	0.07	0.4981	0.9966	
2-4	0.4867	0.4496	0.0282	Да
2-5	0.44	0.4696	0.0763	
2-6	0.345	0.4981	0.3104	
3-4	0.4167	0.4793	0.1172	
3-5	0.37	0.4981	0.241	
3-6	0.275	0.525	0.6117	
4-5	0.04667	0.4496	0.9986	
4-6	0.1417	0.4793	0.9432	
5-6	0.095	0.4981	0.9899	
x2, x3-x1, x5, x6:	0.3583	0.3548	0.0413	Да

**Выводы:** Однофакторный дисперсионный анализ показывает, что сплавы в целом различны по фактору прочности на уровне значимости 0,004. Последующий анализ парных сравнений Шеффе выявляет значимые отличия сплава 2 от стандарта. Заметные отличия от стандарта можно наблюдать и у сплава 3. Тем самым эти два сплава, видимо, имеют структуру, отличную от группы сплавов 1, 5, 6. Дополнительное сравнение двух групп сплавов выявляет значимые различия между ними.

## 8.2.2. Непараметрические методы Крускала-Уоллиса и Джонкхриера

### Пример

**Задача.** В эксперименте измерялось время выполнения монотонных производственных операций в зависимости от мотивационного фактора. Было сформировано три группы рабочих: группа 1 — контрольная; рабочие группы 2 получали только общие сведения о требуемой производительности, а рабочие группы 3 получили полную информацию, включая карты пооперационной разбивки работы (табл. 8.2.3, файл JON). Требуется оценить, влияет ли полученная информация о производительности операций на саму производительность труда.

Таблица 8.2.3. Время выполнения монотонных операций [с] в зависимости от полноты предварительно получаемой информации

	Контроль	Общая информация	Полная информация
Рабочие (разные)	40	38	48
	35	40	40
	38	47	45
	43	44	43
	44	40	46
	41	42	44

### Результаты:

1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: jon.std

Краскал-Уоллис=4.361, Значимость=0.1129, степ.своб = 2

Гипотеза 0: <Нет влияния фактора на отклик>

Джонкхиер=79, Значимость=0.0216, степ.своб = 3,18

Гипотеза 1: <Есть влияние фактора на отклик>

**Выводы:** Хотя критерий Крускала-Уоллиса не выявляет влияния основного фактора, но более конкретизированный анализ с использованием критерия Джонкхиера выявляет значимое влияние уровня чистой мотивации на производительность труда.

**Примечание:** Точные таблицы распределения Крускала-Уоллиса дают для этого примера уровень значимости 0,139 вместо асимптотического 0,1129

### 8.2.3. Непараметрические методы Фридмана и Пейджа

#### Пример 1

**Задача.** В эксперименте измерялось среднее время выточки детали на станке. Работу выполняли 22 рабочих попеременно с использованием трех различных методов обработки (табл. 8.2.4, в файле FREA представлены три переменные *Метод1*, *Метод2*, *Метод3* с 22 измерениями каждая). Требуется оценить, различаются ли эти три метода в плане влияния на производительность работы.

Таблица 8.2.4. Время выточки детали [мин]  
при использовании трех методов работы

Рабочий	1	2	3	4	5	6	7	8	9	10	11
Метод1	5,4	5,85	5,2	5,55	5,9	5,45	5,4	5,45	5,25	5,85	5,25
Метод2	5,5	5,7	5,6	5,5	5,85	5,55	5,4	5,5	5,15	5,8	5,2
Метод3	5,55	5,75	5,5	5,4	5,7	5,6	5,35	5,35	5	5,7	5,1

Рабочий	12	13	14	15	16	17	18	19	20	21	22
Метод1	5,65	5,6	5,05	5,5	5,45	5,55	5,45	5,5	5,65	5,7	6,3
Метод2	5,55	5,35	5	5,5	5,55	5,55	5,5	5,45	5,6	5,65	6,3
Метод3	5,45	5,45	5,95	5,4	5,5	5,35	5,55	5,25	5,4	5,55	6,26

#### Результаты:

2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: frea.std

Фридман=11.14, Значимость=0.0038, степ.своб = 2

Гипотеза 1: <Есть влияние фактора на отклик>

**Выводы:** Результат анализа позволяет принять гипотезу о различии трех методов обработки по влиянию на производительность труда на уровне значимости 0,0038.

#### Пример 2

**Задача.** В эксперименте измерялась прочность хлопка в зависимости от количества калийного удобрения, внесенного в почву при его выращивании. Исследование проводилось на трех различных полях, на каждом из которых выделено пять делянок для испытаний пяти возрастающих доз внесения калийного удобрения (табл. 8.2.5, файл PAGE).

Таблица 8.2.5. Прочность хлопка в зависимости от дозы внесения в почву калийного удобрения

	Доза 1	Доза 2	Доза 3	Доза 4	Доза 5
--	--------	--------	--------	--------	--------

Поле 1	7,46	7,17	7,76	8,14	7,62
Поле 2	7,68	7,57	7,73	8,15	8
Поле 3	7,21	7,80	7,74	7,87	7,93

### **Результаты:**

2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: page.std

Фридман=8.8, Значимость=0.0662, степ.своб = 4

Гипотеза 0: <Нет влияния фактора на отклик>

Пейдж=158, Значимость=0.0039, степ.своб = 5,3

Гипотеза 1: <Есть влияние фактора на отклик>

**В ы в о д ы:** Хотя критерий Фридмана не выявляет влияния основного фактора, но более конкретизированный анализ с использованием критерия Пейджа выявляет значимое увеличение прочности хлопка в зависимости от количества вносимого в почву калийного удобрения на уровне значимости 0,039.

**Примечание:** Точные таблицы распределения Фридмана дают для этого примера уровень значимости 0,0725 вместо асимптотического 0,0662.

## **8.3. Двухфакторный дисперсионный анализ**

### **Пример 1**

**Задача.** В эксперименте фиксировалось урожайность пяти сортов картофеля, выращенных на пяти участках одинакового размера и почвенного состава, при этом каждый из этих участков обрабатывался одним из шести сортов удобрений (табл. 8.3.1, файл A2).

Необходимо выяснить, различна ли в среднем урожайность разных сортов картофеля независимо от применяемого удобрения и различна ли эффективность используемых удобрений независимо от сорта.

*Таблица 8.3.1. Урожайность пяти сортов картофеля [ц/га] в зависимости от вида вносимого в почву удобрения*

	Карт.1	Карт.2	Карт.3	Карт.4	Карт.5
Удобр.1	6	9	6	2	6
Удобр.2	4	7	8	3	5
Удобр.3	9	3	10	7	4
Удобр.4	8	4	14	4	10
Удобр.5	15	11	13	9	14

Удобр.6	12	14	15	11	9
---------	----	----	----	----	---

### Результаты:

2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: a2.std

Факторный план: неповторяемый

Источник	Сум.квадр	Ст.своб	Ср.квадр	Сила влияния
факт.1	79.2	4	19.8	-0.3788
факт.2	248	5	49.6	0.6308
Остат.	118	20	5.9	
Общая	445.2	29	15.35	

F(фактор1)=3.356, Значимость=0.0291, степ.своб = 4,20

Гипотеза 1: <Есть влияние фактора на отклик>

F(фактор2)=8.407, Значимость=0.0003, степ.своб = 5,20

Гипотеза 1: <Есть влияние фактора на отклик>

**Выводы:** Дисперсионный анализ выявляет существенное влияние сорта картофеля и вида удобрения на урожайность (уровни значимости 0,0291 и 0,0003 меньше 0,05).

### Пример 2

**Задача.** В эксперименте измерялось количество выдыхаемого азота для четырех режимов питания и для двух возрастных категорий пациентов. В исследовании участвовало по три разных пациента в каждой из восьми групп «возраст–диета» (табл. 8.3.2, в файле A2G имеется восемь столбцов с тремя измерениями: для диет 1–4 возраста 1 и диет 1–4 возраста 2). Необходимо определить влияние диет, возрастных групп и их взаимодействия на количество выдыхаемого азота.

Таблица 8.3.2. Объем выдыхаемого азота (в литрах) при четырех диетах питания и для двух возрастных групп пациентов

	Диета 1	Диета2	Диета3	Диета4
Возраст 1	4,097	4,368	4,169	4,928
	4,859	5,668	5,709	5,608
	3,54	3,752	4,416	4,94
Возраст 2	2,87	3,579	4,403	4,905
	4,648	5,393	4,496	5,208
	3,848	4,374	4,688	4,806

С иллюстративными целями ниже производятся два расчета для модели с фиксированными эффектами и для модели со случайными эффектами.

### Результаты при фиксированных эффектах:

2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: a2g.std

Факторный план: повторяемый с фиксированными эффектами

Источник	Сум.квадр	Ст.своб	Ср.квадр	Сила влияния
Факт.1	3.63	3	1.21	-0.216
Факт.2	0.335	1	0.335	-0.332
Межфак	0.0453	3	0.0151	-0.333
Остат.	7.83	16	0.489	
Общая	11.8	23	0.515	

F(фактор1)=2.47, Значимость=0.0984, степ.своб = 3,16

Гипотеза 0: <Нет влияния фактора на отклик>  
 $F(\text{фактор2})=0.685$ , Значимость=0.51, степ.своб = 16  
 Гипотеза 0: <Нет влияния фактора на отклик>  
 $F(\text{межфакт})=0.0308$ , Значимость=0.992, степ.своб = 3,16  
 Гипотеза 0: <Нет влияния фактора на отклик>  
 После объединения межфакторной и остаточной СК:  
 $F(\text{фактор1})=2.92$ , Значимость=0.06, степ.своб = 3,19  
 Гипотеза 0: <Нет влияния фактора на отклик>  
 $F(\text{фактор2})=0.808$ , Значимость=0.566, степ.своб = 19  
 Гипотеза 0: <Нет влияния фактора на отклик>

Параметры модели:  
 Среднее = 20.9, доверит.инт.=1.8  
 Эффект1-1 = -16.2, доверит.инт.=14.3  
 Эффект1-1 = -16.5, доверит.инт.=14.3  
 Эффект2-1 = -16.9, доверит.инт.=4.02  
 Эффект2-2 = -16.4, доверит.инт.=4.02  
 Эффект2-3 = -16.2, доверит.инт.=4.02  
 Эффект2-4 = -15.8, доверит.инт.=4.02

### Результаты при случайных эффектах:

2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: a2g.std  
 Факторный план: повторяемый, со случайными эффектами

Источник	Сум.квадр	Ст.своб	Ср.квадр	Сила влияния
Факт.1	3.63	3	1.21	-0.216
Факт.2	0.335	1	0.335	-0.332
Межфак	0.0453	3	0.0151	-0.333
Остат.	7.83	16	0.489	
Общая	11.8	23	0.515	

$F(\text{фактор1})=80.1$ , Значимость=0.0023, степ.своб = 3,3  
 Гипотеза 1: <Есть влияние фактора на отклик>  
 $F(\text{фактор2})=22.2$ , Значимость=0.0002, степ.своб = 3  
 Гипотеза 1: <Есть влияние фактора на отклик>  
 $F(\text{межфакт})=0.0308$ , Значимость=0.992, степ.своб = 3,16  
 Гипотеза 0: <Нет влияния фактора на отклик>  
 После объединения межфакторной и остаточной СК:  
 $F(\text{фактор1})=2.92$ , Значимость=0.06, степ.своб = 3,19  
 Гипотеза 0: <Нет влияния фактора на отклик>  
 $F(\text{фактор2})=0.808$ , Значимость=0.566, степ.своб = 19  
 Гипотеза 0: <Нет влияния фактора на отклик>

**В ы в о д ы:** Для модели с фиксированными эффектами не обнаружено значимых эффектов факторов или их взаимодействия. Для модели со случайными эффектами оба фактора оказались значимыми. Поскольку же их взаимодействие незначимо, то исследователю может показаться целесообразным произвести объединение, так как число степеней свободы остаточной дисперсии (знаменателя F-критерия) слишком мало для оценки эффектов. Однако в результате объединения оба фактора становятся незначимыми. Эти изменения в оценке значимости показывают, насколько различными могут оказаться результаты в зависимости от отношения исследователя к типу модели и объединению дисперсий.

## 8.4. Дисперсионный анализ групповых измерений

### Пример 1

**Задача.** Рассмотрим однофакторный эксперимент, в котором девять пациентов используют релаксационную методику для ослабления мигрирующей головной боли (фиксируется число приступов в неделю). Исследование проводилось в течение пяти недель, причем релаксационная техника использовалась только в последние две недели. Тем самым стоит задача выявить влияние фактора релаксации на фоне первых двух контрольных недель. Полученные данные приведены в табл. 8.4.1 (файл ANOVA\_BS, где переменные (столбцы) соответствуют неделям, а строки — пациентам).

Таблица 8.4.1. Частота головных болей при релаксационной терапии

Пациенты	Недели				
	1	2	3	4	5
1	21	22	8	6	6
2	20	19	10	4	4
3	17	15	5	4	5
4	25	30	13	12	17
5	30	27	13	8	6
6	19	27	8	7	4
7	26	16	5	2	5
8	17	18	8	1	5
9	26	24	14	8	9

В меню рис. 8.8 выбираем «1-факторный» групповой анализ.

### Результаты:

```

ГРУППОВОЙ 1-ФАКТОРНЫЙ АНАЛИЗ.  Файл: anova_bs.std
Источник  Сум.квадр  Ст.своб  Ср.квадр
Групповая      486.7      8
Факт.1         2449      4      612.3
Остаточн.      230.4     32      7.2
Общая         3166     44
F(фактор1)=85.04,  Значимость=0,  степ.своб = 4,32
Гипотеза 1: <Есть влияние фактора на отклик>
Матрица ковариаций
      21      11.75      9.25      7.833      7.333
      11.75      28.5      13.75      16.38      13.38
      9.25      13.75      11.5      8.583      8.208
      7.833      16.38      8.583      11.69      10.82
      7.333      13.38      8.208      10.82      16.94
Поправочные коэффициенты: e=0.6845  e'=1
С учетом поправочного коэффициента=0.6845
F(фактор1)=85.04,  Значимость=0,  степ.своб = 2.738,21.9
Гипотеза 1: <Есть влияние фактора на отклик>

```



**Выводы:** Результаты проведенного анализа выявляют влияние фактора релаксационной тренировки (уровень значимости близок к нулю). Отметим, что если бы мы использовали здесь обычную технику однофакторного анализа (применимую и в случае пяти различных групп пациентов в каждой из пяти недель и поэтому приводящую для нашего случая к завышению остаточной дисперсии), то получили бы значение критерия Фишера, равное 34,15 с 40 степенями свободы, которое хотя и тоже значимо, но в 2 с лишним раза меньше вышеприведенного.

## Пример 2

**Задача.** Рассмотрим двухфакторный эксперимент с одним межгрупповым фактором, в котором исследуется двигательная активность крыс в ответ на инъекции мидозалама (King, 1986). В первое время после инъекции обычно наблюдается снижение моторики, однако, как и в случае морфина, здесь быстро развивается привыкание. Кинг попытался выяснить, можно ли объяснить степень этого привыкания только внешними условиями. В предэкспериментальной фазе двум тестовым группам (из восьми животных каждая) вводились инъекции препарата в течение нескольких дней, а контрольной группе из восьми животных вводился нейтральный физиологический раствор. В заключительный (тестовый) день всем трем группам были сделаны инъекции мидозалама, причем одна из тестовых групп находилась в своем обычном помещении (среде), а другая тестовая группа была перенесена в новую окружающую среду. Если априорная гипотеза Кинга верна, то реакции контрольной группы должны быть близки к реакциям группы, сменившей среду обитания, поскольку в обоих этих группах исключался фактор условного привыкания, определенный неизменностью среды обитания. Поскольку процесс метаболизма лекарства составляет около одного часа, измерения двигательной активности проводились в шести последовательных 5-минутных интервалах, и собранные данные приведены в табл. 8.4.2 (файл ANOVA\_B2, где переменные (столбцы) соответствуют временным интервалам для трех последовательных групп, а строки — животным).

Таблица 8.4.2. Двигательная активность крыс в ответ на инъекцию мидозалама для шести временных интервалов и трех экспериментальных групп

Группы	Временные интервалы					
	1	2	3	4	5	6
1. Контроль	150	44	71	59	132	74
	335	270	156	160	118	230
	149	52	91	115	43	154
	159	31	127	212	71	224
	159	0	35	75	71	34
	292	125	184	246	225	170
	297	187	66	96	209	74

	170	37	42	66	114	81
2. Та же среда	346	175	177	192	239	140
	426	329	236	76	102	232
	359	238	183	123	183	30
	272	60	82	85	101	98
	200	271	263	216	241	227
	366	291	263	144	220	180
	371	364	270	308	219	267
	497	402	294	216	284	255
3. Новая среда	282	186	225	134	189	169
	317	31	85	120	131	205
	362	104	144	114	115	127
	338	132	91	77	108	169
	263	94	141	142	120	195
	138	38	16	95	39	55
	329	62	62	6	93	67
	292	139	104	184	193	122

Очевидно, что межгрупповой эффект суммируется из различия между группами и различия между объектами в каждой группе, в то время как внутригрупповой (временной) эффект имеет три компонента: главный эффект повторно–временных измерений и его два взаимодействия с групповыми различиями и с различиями объектов.

В меню (рис. 8.8) выбираем: 1–меж– и 1–внутригрупповой анализ, устанавливая в этом меню шесть градаций первого фактора.

### Результаты:

ГРУППОВОЙ 1+1–ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.

Файл: anova\_b2.std

Источник	Сум.квадр	Ст.своб	Ср.квадр
Групповая	6.705E5	23	
Факт.1	3.997E5	5	7.995E4
Межфакт.	8.082E4	10	8082
Остат.1	2.812E5	105	2678
Факт.2	2.858E5	2	1.429E5
Остат.2	3.847E5	21	1.832E4
Общая	1.432E6	143	

F(фактор1)=29.85, Значимость=0, степ.своб = 5,70

Гипотеза 1: <Есть влияние фактора на отклик>

F(фактор2)=7.801, Значимость=0.0055, степ.своб = 2,14

Гипотеза 1: <Есть влияние фактора на отклик>

F(межфактор.)=3.018, Значимость=0.0034, степ.своб = 10,70

Гипотеза 1: <Есть влияние фактора на отклик>

Матрица ковариаций

6388	4696	2240	681.6	2018	1924
4696	7864	4181	2462	2892	3532
2240	4181	3912	2697	2162	3298
681.6	2462	2697	4601	2249	3085
2018	2892	2162	2249	3717	989.3
1924	3532	3298	3085	989.3	5228

Поправочные коэффициенты: e=0.6569 e'=0.8674

С учетом поправочного коэффициента=0.6569

F(фактор1)=29.85, Значимость=0, степ.своб = 3.285,45.99

Гипотеза 1: <Есть влияние фактора на отклик>

F(фактор2)=7.801, Значимость=0.0055, степ.своб = 2,14

Гипотеза 1: <Есть влияние фактора на отклик>  
 $F(\text{межфактор.}) = 3.018$ ,  $\text{Значимость} = 0.0121$ ,  $\text{степ.своб} = 6.569, 45.99$   
 Гипотеза 1: <Есть влияние фактора на отклик>

**В ы в о д ы:** Результаты проведенного анализа выявляют существенное влияние как внутригруппового (временного) фактора, так и межгрупповые различия от внешних условий, а также и их взаимодействие (уровни значимости 0, 0,0055 и 0,0034 меньше 0,05). На графике факторных эффектов (рис. 8.8) видно, что активность животных существенно падает в первом 5–минутном интервале, причем ее уровень в неизменной среде превышает активность других двух групп, но падение активности продолжается в следующих интервалах, тогда как в других группах она немного повышается.

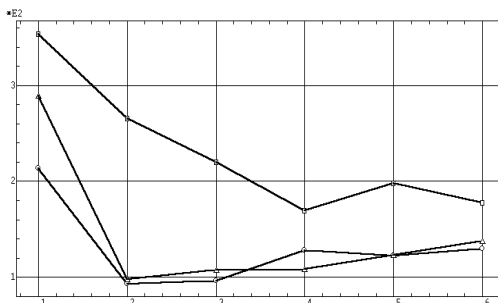


Рис. 8.9. График факторных эффектов. По горизонтальной оси — уровни первого фактора, порядок графиков (сверху вниз): та же среда, контроль, новая среда

Последующий анализ может идти по пути исследования *простых эффектов* посредством сведения двухфакторной схемы к различным вариантам однофакторной, например, для исследования различий между группами в первом и последнем временном интервале (классический однофакторный анализ с повторными измерениями), или для исследования влияния фактора времени в каждой из трех групп по отдельности (однофакторный групповой анализ).

**З а д а ч а.** Модифицируем теперь условно схему данного эксперимента, считая что во всех группах были одни и те же крысы (т. е. как бы имея план с двумя внутригрупповыми факторами).

### Результаты:

ГРУППОВОЙ 2-ФАКТОРНЫЙ АНАЛИЗ. Файл: anova\_b2.std  
 Источник Сум.кв.адр Ст.своб Ср.кв.адр  
 Групповая 6.378E4 7  
 Факт.1 3.997E5 5 7.995E4  
 Остат.1 1.018E5 35 1.72E4  
 Факт.2 2.858E5 2 1.429E5  
 Остат.2 3.209E5 14 2.292E4  
 Межфакт12 8.082E4 10 8082  
 Остат.12 6.021E5 70 8602

Общая 1.432Е6 143  
 F(фактор1)=27.49, Значимость=0, степ.своб = 5,35  
 Гипотеза 1: <Есть влияние фактора на отклик>  
 F(фактор2)=6.234, Значимость=0.0115, степ.своб = 2,14  
 Гипотеза 0: <Есть влияние фактора на отклик>  
 F(межфактор.12)=0.9395, Значимость=0.5036, степ.своб = 10,70  
 Гипотеза 0: <Нет влияния фактора на отклик>  
 Матрица ковариаций  

6388	4696	2240	681.6	2018	1924
4696	7864	4181	2462	2892	3532
2240	4181	3912	2697	2162	3298
681.6	2462	2697	4601	2249	3085
2018	2892	2162	2249	3717	989.3
1924	3532	3298	3085	989.3	5228

 Поправочные коэффициенты: e=0.6569 e''=0.8674  
 С учетом поправочного коэффициента=0.6569  
 F(фактор1)=27.49, Значимость=0, степ.своб = 3.285,45.99  
 Гипотеза 1: <Есть влияние фактора на отклик>  
 F(фактор2)=6.234, Значимость=0.0115, степ.своб = 2,14  
 Гипотеза 1: <Есть влияние фактора на отклик>  
 F(межфактор.)=0.9395, Значимость=0.5171, степ.своб=6.569,45.99  
 Гипотеза 0: <Нет влияния фактора на отклик>

**В ы в о д ы:** Отличие полученных результатов от вышерассмотренной схемы с одним внутригрупповым и одним межгрупповым факторами состоит в резком усилении незначимости межфакторного взаимодействия и в двукратном увеличении уровня значимости нулевой гипотезы для второго фактора.

## Пример 3

**З а д а ч а.** Рассмотрим трехфакторный эксперимент (схема с одним внутригрупповым фактором) по использованию индивидуальных профилактических приемов для предотвращения угрозы заболевания при контактах с потенциально загрязненной средой (табл. 8.4.3, файл ANOVA\_B3, где строки отвечают испытуемым, а столбцы — последовательным уровням факторов в последовательности: время, методика, пол). Использовались два вида обучения, которые проходили две группы испытуемых: лекционный курс и индивидуальная тренировка выполнения необходимой профилактики. Каждая группа включала 10 мужчин и 10 женщин, у которых по данным опроса выяснялась частота выполнения профилактики до и после обучения, а также спустя 6 месяцев и 1 год.

Таблица 8.4.3. Профилактика от заболевания

	Тренировка				Демонстрация			
	До	После	6 мес	1 год	До	После	6 мес	1 год
Мужчины	7	22	13	14	0	0	0	0
	25	10	17	24	69	56	14	36
	50	36	49	23	5	0	0	5
	16	38	34	24	4	24	0	0
	33	25	24	25	35	8	0	0
	10	7	23	26	7	0	9	37

	13	33	27	24	51	53	8	26
	22	20	21	11	25	0	0	15
	4	0	12	0	59	45	11	16
	17	16	20	10	40	2	33	16
Женщины	0	6	22	26	15	28	26	15
	0	16	12	15	0	0	0	0
	0	8	0	0	6	0	23	0
	15	14	22	8	0	0	0	0
	27	18	24	37	25	28	0	16
	0	0	0	0	36	22	14	48
	4	27	21	3	19	22	29	2
	26	9	9	12	0	0	5	14
	0	0	14	1	0	0	0	0
	0	0	12	0	0	0	0	0

Таким образом, данная схема отвечает трехфакторному плану с одним внутригрупповым фактором (время) и двумя межгрупповыми факторами (методика обучения и пол). В рис. 8.8 выбираем: «1–внутригрупповой» и «2–межгрупповой» анализ, устанавливая в этом меню четыре градации первого фактора и две градации второго фактора

### Результаты:

```

ГРУППОВОЙ 1+2-ФАКТОРНЫЙ АНАЛИЗ.  Файл: anova_b3.std
Источник  Сум.кв.адр  Ст.своб  Ср.кв.адр
Групповая  2.149E4      39
Факт.1     274.1       3      91.36
Межфакт12  1378          3     459.3
Межфакт13  779.9         3     260
Межфакт123 476.4         3     158.8
Остат.1     1.101E4      108    101.9
Факт.2     107.3        1     107.3
Межфакт23  63.76         1     63.76
Факт.3     3358          1     3358
Остат.2     1.796E4      36     498.9
Общая      3.54E4       159
F(фактор1)=0.8965, Значимость=0.5521, степ.своб = 3,108
  Гипотеза 0: <Нет влияния фактора на отклик>
F(фактор2)=0.215, Значимость=0.8254, степ.своб = 36
  Гипотеза 0: <Нет влияния фактора на отклик>
F(фактор3)=6.731, Значимость=0, степ.своб = 36
  Гипотеза 1: <Есть влияние фактора на отклик>
F(межфактор.12)=4.507, Значимость=0.0054, степ.своб = 3,108
  Гипотеза 1: <Есть влияние фактора на отклик>
F(межфактор.13)=2.551, Значимость=0,0583, степ.своб = 3,108
  Гипотеза 0: <Нет влияния фактора на отклик>
F(межфактор.23)=0.1278, Значимость=0.8944, степ.своб = 36
  Гипотеза 0: <Нет влияния фактора на отклик>
F(межфактор.123)=1.558, Значимость=0.2023, степ.своб = 3,12
  Гипотеза 0: <Нет влияния фактора на отклик>
Матрица ковариаций
  335.7      206      74.55     131.6
    206      251      75.25     100
    74.55    75.25     146      64.17
    131.6    100      64.17    168.1

```

Поправочные коэффициенты: e=0.8649 e''=1

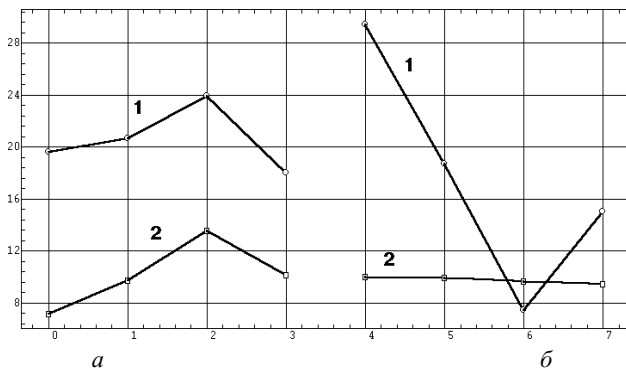


Рис. 8.10. График факторных эффектов (по горизонтальной оси — уровни первого фактора): для тренированной группы (а); для лекционной группы (б); 1 — мужчины; 2 — женщины

**В ы в о д ы:** Как показывают полученные результаты, значимым является только фактор пола (уровень значимости близок к нулю), а также взаимодействие факторов *время–методика* (уровень значимости 0,0054). Изучение графика факторных эффектов (рис. 8.10) показывает, что женщины пользуются профилактическими приемами реже, чем мужчины, однако это может свидетельствовать и о том, что первые просто реже контактируют с потенциально загрязненной средой. Этот же график уточняет тенденцию обнаруженного межфакторного взаимодействия: обе тренируемые группы усиливают профилактические действия с течением времени, тогда как “лекционные” группы не изменяют или даже снижают свой первоначальный уровень. Поскольку поправочные коэффициенты  $\epsilon$  и  $\epsilon'$  велики по своим значениям, повторная проверка гипотез со скорректированными степенями свободы не проводится.

Дальнейший анализ может идти по пути детального исследования *простых эффектов* путем сведения трехфакторной схемы к различным комбинациям двухфакторной: пол и время в тренируемых (“лекционных”) группах, пол и методика при различных временных срезах, методика и время для каждого пола и т. п.

## Пример 4

**З а д а ч а.** Рассмотрим трехфакторный эксперимент (схема с двумя внутригрупповыми факторами) по выработке у крыс условного рефлекса на болевой раздражитель (табл. 8.4.4, файл ANOVA\_B4, где строки отвечают крысам, а столбцы — последовательным уровням факторов в последовательности: фаза, цикл, группа). Нормально крысы научены нажимать на педальки в своем вольере для получения пищи, чем они и занимаются значительную часть времени. Далее одной группе из 8 крыс (А) в фазе I периодически подают звуковой сигнал (предупредительный

стимул), после чего на педальки поступает слабomощный электрический заряд для болевого воздействия, что влечет снижения активности крыс по нажатию на педальку в результате научения.

*Таблица 8.4.4. Выработка условного рефлекса у крыс на предупредительный стимул болевого раздражителя*

	Ц и к л ы							
	1		2		3		4	
	фаза		фаза		фаза		фаза	
Группа	I	II	I	II	I	II	I	II
А	1	28	22	48	22	50	14	48
	21	21	16	40	15	39	11	56
	15	17	13	35	22	45	1	43
	30	34	55	54	37	57	57	68
	11	23	12	33	10	50	8	53

*Продолжение таблицы 8.4.4*

	16	11	18	34	11	40	5	40
	7	26	29	40	25	50	14	56
	0	22	23	45	18	38	15	50
В	1	6	16	8	9	14	11	33
	37	59	28	36	34	32	26	37
	18	43	38	50	39	15	29	18
	1	2	9	8	6	5	5	15
	44	25	28	42	47	46	33	35
	15	14	22	32	16	23	32	26
	0	3	7	17	6	9	10	15
	26	15	31	32	28	22	16	15
С	33	43	40	52	39	52	38	48
	4	35	9	42	4	46	23	51
	32	39	38	47	24	44	16	40
	17	34	21	41	27	50	13	40
	44	52	37	48	33	53	33	43
	12	16	9	39	9	59	13	45
	18	42	3	62	45	49	60	57
	13	29	14	44	9	50	15	48

В фазе II эту группу крыс переводят в другой вольер (изменение окружающей среды), где звуковой сигнал не подкрепляется последующим электрическим шоком. И так проводятся четыре последовательных цикла из этих двух фаз. В этой группе в результате выработанного условного рефлекса частота нажатий на педальку в фазе II также должна быть ниже по сравнению с отсутствием звукового стимула, однако она должна повышаться от цикла к циклу в связи с влиянием фактора “безопасной среды” в фазе II. У второй группы крыс (В) аналогичный рефлекс вырабатывается в фазе I не на звуковой, а на световой стимул, поэтому в новом вольере фазы II они в ответ на звуковой стимул должны менее реагировать на звук, особенно в первых двух циклах. Третья группа крыс (С) отличалась от первой только тем, что обе фазы выполнялись в том же самом вольере, т. е. не присутствовал фактор “безопасной среды”. В

эксперименте фиксировалось число нажатий на педалики после подачи предупредительного стимула.

Таким образом, данная схема отвечает трехфакторному плану с двумя внутригрупповыми факторами (фаза и цикл) и одним межгрупповым фактором (три различные условно–рефлекторные методики). В меню рис. 8.8 выбираем: «2–внутригрупповой» и «1–межгрупповой анализ», устанавливая в этом меню две градации первого фактора и четыре градации второго фактора.

### Результаты:

ГРУППОВОЙ 2+1–ФАКТОРНЫЙ АНАЛИЗ. Файл: anova\_b4.std

Источник	Сум.кв.др	Ст.своб	Ср.кв.др
Групповая	2.034E4	23	
Факт.1	1.17E4	1	1.17E4
Межфакт13	4054	2	2027
Остат.1	1893	21	90.12
Межфакт12	741.5	3	247.2
Межфак123	1274	6	212.3
Остат.1–3	3859	63	61.26
Факт.2	2727	3	909
Межфакт23	1047	6	174.5
Остат.2	4761	63	75.58
Факт.3	4617	2	2308
Остат.3	1.572E4	21	748.7
Общая	5.24E4	191	

F(фактор1)=129.9, Значимость=0, степ.своб = 21

Гипотеза 1: <Есть влияние фактора на отклик>

F(фактор2)=12.03, Значимость=0, степ.своб = 3,63

Гипотеза 1: <Есть влияние фактора на отклик>

F(фактор3)=3.083, Значимость=0.0656, степ.своб = 2,21

Гипотеза 0: <Нет влияния фактора на отклик>

F(межфактор.12)=4.035, Значимость=0.0109, степ.своб = 3,63

Гипотеза 0: <Нет влияния фактора на отклик>

F(межфактор.13)=22.49, Значимость=0, степ.своб = 2,21

Гипотеза 1: <Есть влияние фактора на отклик>

F(межфактор.23)=2.309, Значимость=0.044, степ.своб = 6,63

Гипотеза 0: <Нет влияния фактора на отклик>

F(межфактор.123)=3.466, Значимость=0.0053, степ.своб = 6,63

Гипотеза 1: <Есть влияние фактора на отклик>

**В ы ы в о д ы:** Как показывают полученные результаты, значимыми являются первые два фактора (их уровни значимости близки к нулю), т. е. имеется сильное влияние как фазы (подкрепляемый и неподкрепляемый предупредительный стимул), так и цикла (т. е. временного дообучения тому, что фаза II безопасна). С другой стороны, неожиданно оказывается малозначимым третий фактор группы (уровень значимости 0,0656 выше уровня 0,05), т. е. фактор дополнительной смены внешних условий между фазами или смены модальности предупредительного стимула. Вместе с тем оказываются значимыми все три парных межфакторных взаимодействия (0,0109, 0, 0,044), так же как и трехфакторное взаимодействие (0,0053).



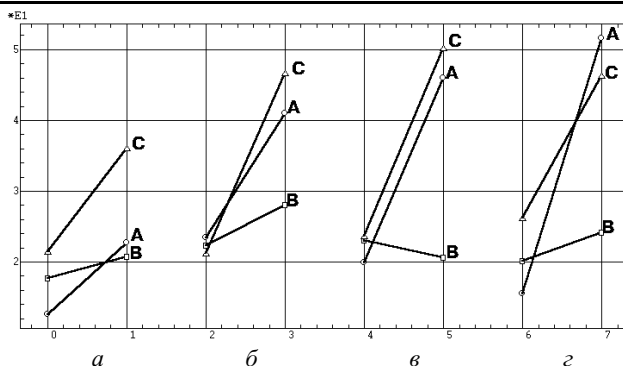


Рис. 8.11. График факторных эффектов с первого по четвертый циклы (а, б, в, г); по горизонтальной оси — уровни первого фактора; на графиках по вертикали указаны обозначения групп А, В, С

Кроме того, рисунок факторных эффектов (рис. 8.11) показывает, что первая группа демонстрирует значительно меньшее обучение “безопасному предупреждающему стимулу” в фазе II, т. е. смена внешних условий оказывает дополнительное пролонгирующее шоковое ожидание. Поэтому, если исключить из анализа одну из групп В или С, то и эффект третьего фактора окажется значимым. Аналогичным образом, исключая ту или иную фазу или выбирая отдельные циклы (т. е. сводя трехфакторную схему к двухфакторной), можно изучать более тонкие взаимодействия между исследуемыми факторами и группами.

## Пример 5

**З а д а ч а.** Рассмотрим схему трехфакторно–внутригруппового эксперимента (табл. 8.4.5, файл ANOVA\_B5, где строки отвечают водителям, а столбцы — последовательным уровням факторов в последовательности: модель, время суток, дорога): имеются три автомобиля различных моделей: дешевая, средняя, дорогая (М, С, Д), три типа дорожного покрытия (грунтовая, щебенка и шоссе) и два времени суток (день и ночь). Три водителя управляют каждым из трех автомобилей во всех комбинациях условий и у них фиксируется число ошибок вождения.

Таблица 8.4.5. Число ошибок вождения у трех водителей в зависимости от модели автомобиля, времени суток и дорожного покрытия

	Ночь			День		
	М	С	Д	М	С	Д
Дорога:						
Грунтовая	10	8	6	5	4	3
	9	8	5	4	3	3
	8	7	4	4	1	2

Щебенка	9	7	5	4	3	3
	10	6	4	4	2	2
	7	4	3	3	3	2
Шоссе	7	6	3	2	2	1
	4	5	2	2	3	2
	3	4	2	1	0	1

В меню рис. 8.8 выбираем: «3–внутригрупповой» анализ, устанавливая в этом меню три градации первого фактора и две градации второго фактора.

### Результаты:

ГРУППОВОЙ 3–ФАКТОРНЫЙ АНАЛИЗ. Файл: anova\_b5.std

Источник	Сум.квадр	Ст.своб	Ср.квадр
Групповая	24.11	2	
Факт.1	51.44	2	25.72
Остат.1	1.111	4	0.2222
Факт.2	140.2	1	140.2
Остат.2	2.333	2	1.167
Факт.3	56.78	2	28.39
Остат.3	0.1111	4	0.02778
Межфакт12	16.78	2	8.389
Остат.12	0.8889	4	0.2222
Межфакт13	8.778	4	2.194
Остат.13	4.667	8	0.5833
Межфакт23	5.444	2	2.722
Остат.23	5.222	4	1.306
Межфак123	2.778	4	0.6944
Остат.123	2.889	8	0.3611
Общая	323.5	53	

F(фактор1)=92.6, Значимость=0, степ.своб = 2,4

Гипотеза 1: <Есть влияние фактора на отклик>

F(фактор2)=120.1, Значимость=0.0002, степ.своб = 2

Гипотеза 1: <Есть влияние фактора на отклик>

F(фактор3)=1022, Значимость=0.0002, степ.своб = 2,4

Гипотеза 1: <Есть влияние фактора на отклик>

F(межфактор.12)=37.75, Значимость=0.0041, степ.своб = 2,4

Гипотеза 1: <Есть влияние фактора на отклик>

F(межфактор.13)=3.762, Значимость=0,0524, степ.своб = 4,8

Гипотеза 0: <Нет влияния фактора на отклик>

F(межфактор.23)=2.085, Значимость=0.2396, степ.своб = 2,4

Гипотеза 0: <Нет влияния фактора на отклик>

F(межфактор.123)=1.923, Значимость=0.1997, степ.своб = 4,8

Гипотеза 0: <Нет влияния фактора на отклик>

**В ы в о д ы:** Как показывают полученные результаты, на число ошибок вождения оказывают влияние все три исследуемых фактора (их уровни значимости близки к нулю), а также наблюдается значимое взаимодействие факторов модели автомобиля и времени суток (уровень значимости 0.0041 существенно ниже уровня 0,05).

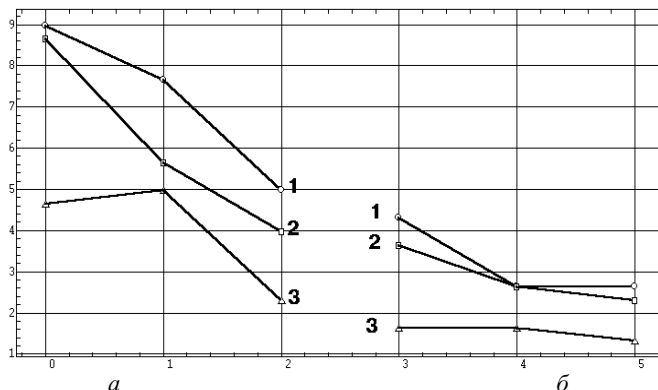


Рис. 8.12. График факторных эффектов (по горизонтальной оси — уровни первого фактора) для ночи и для дня (а, б);  
1 — дешевые автомобили; 2 — средние; 3 — дорогие

Как показывает график факторных эффектов (рис. 8.12), число ошибок вождения существенно падает при улучшении дорожного покрытия, при переходе к более дорогому автомобилю и в дневное время суток по сравнению с ночным.

## 8.5. Многофакторный дисперсионный анализ

### Пример

**Задача.** Исследовалось влияние трех агротехнических технологий выращивания картофеля и двух методов предпосевной обработки семян на урожайность. Для каждого сочетания значений этих двух факторов было выделено разное число участков. Полученные данные об урожайности картофеля представлены в табл. 8.5.1.

Таблица 8.5.1. Урожайность картофеля [ц/га] в зависимости от предпосевной обработки и агротехнических методов выращивания

Фактор $F_2$ — предпосевная обработка	Фактор $F_1$ — агротехнический метод		
	1	2	3
1	17.5 16	13.2	12.8 10.4 9.9
2	10.1 8.6 11.3	5.4 3.7	10.3

Такого типа данные в	F <sub>1</sub>	F <sub>2</sub>	Y
электронной таблице для	1	1	17.5
анализа представляются в виде	1	1	16.2
матрицы, содержащей три	2	1	13.2
переменные: значения первого	3	1	12.8
и второго факторов и значение	3	1	10.4
измеренного отклика (файл	3	1	9.9
MAV):	1	2	10.1
	1	2	8.6
	1	2	11.3
	2	2	5.4
	2	2	3.7
	3	2	10.3

### Результаты:

МНОГОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: mav.std

Источник Сум.квадр Ст.своб Ср.квадр Сила влияния  
 Фактор 1 92.15 2 46.07 0.1648  
 F(фактор 1)=6.073, Значимость=0.0247, степ.своб = 2,8  
 Гипотеза 1: <Есть влияние фактора на отклик>  
 Фактор 2 117.3 1 117.3 0.663  
 F(фактор 2)=17.66, Значимость=0, степ.своб = 8  
 Гипотеза 1: <Есть влияние фактора на отклик>  
 Остат. 36.59 8 4.574

**В ы в о д ы:** Результаты анализа показывают значимое влияние на урожайность как предпосевной обработки семян, так и агротехнической технологии.

**З а д а ч а 2:** Эти же данные представляется возможным обработать методом двухфакторного анализа с повторными измерениями (см. разд. 8.3, в соответствии с требованиями метода данные представлены в файле MAV1) и сравнить результаты.

### Результаты:

2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ. Файл: mav1.std

Факторный план: повторяемый, с фиксированными эффектами

Источник	Сум.квадр	Ст.своб	Ср.квадр	F	Значимость	Сила влияния
Факт.1	39.8	2	19.9	10.8	0.0109	0.429
Факт.2	84.1	1	84.1	45.7	1.67E-5	0.483
Межфак	33.1	2	16.6	9	0.0161	0.4
Остат.	11	6	1.84			
Общая	168	11	15.3			

F(фактор1)=10.8, Значимость=0.0109, степ.своб = 2,6  
 Гипотеза 1: <Есть влияние фактора на отклик>  
 F(фактор2)=45.7, Значимость=1.67E-5, степ.своб = 6  
 Гипотеза 1: <Есть влияние фактора на отклик>  
 F(межфакт)=9, Значимость=0.0161, степ.своб = 2,6  
 Гипотеза 1: <Есть влияние фактора на отклик>

**В ы в о д ы:** При сравнении этих результатов с предыдущими можно заметить некоторые расхождения в суммах квадратов и степенях свободы, поскольку многофакторный метод не учитывает межфакторных взаимодействий. Тем не менее результирующие значимости нулевых гипотез находятся в очень хорошем согласии.

## 8.6. Ковариационный анализ

### Пример

**З а д а ч а.** В эксперименте изучалось влияние тренировки на способность человека близко подойти к устрашающему объекту (живой змее), прежде чем он почувствует дискомфорт или беспокойство. Набрали 4 группы по 10 добровольцев (всего 40 испытуемых), с которыми провели и различные по используемому манекену змеи предварительные тренировки. Эти четыре группы соответствовали четырем уровням фактора тренировки  $F$ , влияние которого следовало выяснить в результате анализа.

Далее провели реальные испытания, фиксируя минимальное расстояние приближения к живой змее (переменная  $Y$ ). Однако это расстояние может зависеть от множества других сопутствующих факторов (смелость, возраст, острота зрения и пр.), поэтому при неудачном разбиении испытуемых на группы статистический результат может быть сильно искажен. В данном случае решили учитывать один сопутствующий фактор — природную смелость, оцениваемую аналогичной пробой до начала тренировки (переменная  $X$ ). Данные исследования представлены в табл. 8.6.1, а для анализа эти данные представлены (файл COV) значениями трех парных переменных:  $X$  — все значения до тренировки;  $Y$  — все значения после тренировки;  $F$  — соответствующие значения фактора.

Таблица 8.6.1. Расстояние приближения к живой змее (в дюймах)

F-тип манекена	1	2	3	4	1	2	3	4
Испытуемые	X-до тренировки				Y-после тренировки			
1	25	17	32	10	25	11	24	8
2	13	9	30	29	25	9	18	17
3	10	19	12	7	12	16	2	8
4	25	25	30	17	30	17	24	12
5	10	6	10	8	37	1	2	7
6	17	23	8	30	25	12	0	26
7	9	7	5	5	31	4	0	8
8	18	5	11	29	26	3	1	29
9	27	30	5	5	28	26	1	29
10	17	19	25	13	29	20	10	9

### Результаты:

КОВАРИАЦИОННЫЙ АНАЛИЗ. Файл: cov.std

Гомогенность X1:  $F=0.07494$ , Значимость= $0.9723$ , степ.своб = 3,36

Гипотеза 0: <Нет влияния фактора на отклик>

Гомогенность X2:  $F=8.85$ , Значимость= $0.0003$ , степ.своб = 3,36

Гипотеза 1: <Есть влияние фактора на отклик>

Коэфф.	a0	a1			
Значение	4.262E-12	0.6427			
Ст.ошиб.	0.9321	0.1047			
Значим.	0.9955	0			
Источник	Сум.квадр.	Степ.св	Средн.квадр.		
Регресс.	1310	1	1310		
Остаточн	1321	38	34.75		
Вся	2630	39			
Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим
0.7056	0.4979	0.4847	5.895	37.68	0
Гипотеза 1: <Есть влияние фактора на отклик>					
Источник	Сум.квадр	Ст.своб	Ср.квадр	Сила влияния	
Факт.1	1940	3	646.6	0.3666	
Остат.	1321	35	37.73		
Общая	3260	38	85.8		
F(фактор1)=17.63, Значимость=0, степ.своб = 3,36					
Гипотеза 1: <Есть влияние фактора на отклик>					

**В ы в о д ы:** Предварительные результаты показывают, что значения сопутствующей переменной достаточно однородно распределены по уровням фактора, а значения переменной–отклика существенно различаются в этих же группах. Регрессионная модель также достаточно хорошо воспроизводит зависимость отклика от сопутствующей переменной. Заключительный дисперсионный анализ позволяет отклонить гипотезу об отсутствии влияния фактора на уровне значимости, близком к нулю, и принять гипотезу о присутствии фактора тренировки.

## АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

### 9.2. Корреляционный анализ

#### Примеры

**З а д а ч а.** Некая коммерческая фирма уделяла важное внимание работе на рынке фьючерсных контрактов. Поэтому для выработки обоснованной стратегии желательно понимание динамики фьючерсного курса и его несомненной связи с текущим курсом доллара.

Для этого используются (табл. 9.2.1) результаты долларовых торгов на ММВБ в период с 3.10.94 по 15.12.94 (переменная *dol* из файла SPEC) и результаты фьючерсных долларовых торгов с поставкой 15.12.94 (торги производятся каждый день, исключая субботу и воскресенье, переменная *fut*)

Таблица 9.2.1. Результаты долларовых и фьючерсных торгов на ММВБ в период с 3.10.94 по 15.12.94

доллар	2643, 2668, 2808, 2833, 3081, 3926, 2994, 2988, 2996,
--------	---

	2996
фьючерс	3104, 3228, 3386, 3524, 3684, 3524, 3524, 3585, 3582, 3590
доллар	3005, 3015, 3030, 3036, 3046, 3055, 3075, 3085, 3093, 3099
фьючерс	3626, 3550, 3485, 3484, 3487, 3476, 3478, 3485, 3482, 3458
доллар	3102, 3102, 3102, 3118, 3131, 3143, 3157, 3175, 3187, 3198
фьючерс	3368, 3303, 3281, 3315, 3358, 3403, 3379, 3379, 3370, 3370
доллар	3201, 3228, 3232, 3234, 3249, 3275, 3275, 3292, 3306, 3338
фьючерс	3380, 3376, 3377, 3355, 3350, 3362, 3355, 3350, 3359, 3380
доллар	3368, 3383
фьючерс	3380, 3380

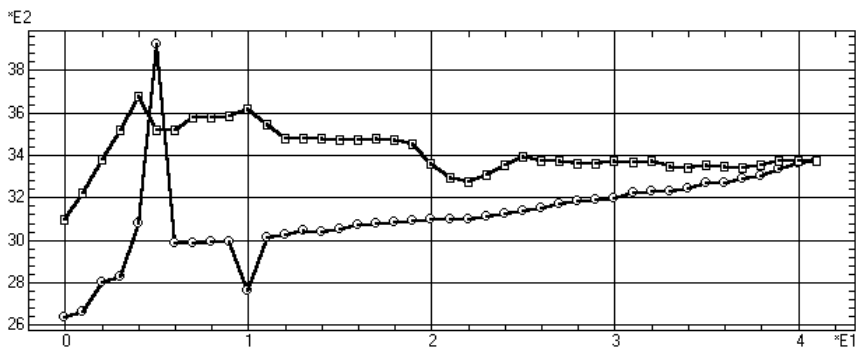


Рис. 9.2. Котировка доллара на ММВБ (кресты) и фьючерсных поставок (квадраты) с 3.10 по 15.12 1994 г.

**Визуальный анализ.** В курсе доллара (рис. 9.2) выделяется пик знаменитого «черного вторника» (11.10.94) и линейная тенденция с небольшими колебаниями, отражающая методику «циркуля и линейки», распространенную тогда в высших планирующих органах. Во фьючерсном курсе имеется вполне понятная тенденция приближения к реальному курсу (при приближении дня поставки 15.12.94) с нерегулярными и достаточно высокоамплитудными колебаниями. Интересно, что курс фьючерсных контрактов рос при начальном повышении курса доллара, но заметно не отреагировал собственно на *черный вторник*. Тем самым, уже простое изучение графиков временных рядов дает нам достаточно много предварительной информации.

**Постановка задачи.** Очевидно, что упомянутые процессы развиваются в некоторой информационной среде, обладающей собственной инерционностью и упругостью, что определено существующими

организационной и инфраструктурой, менталитетом участников и заинтересованных лиц и кучей других факторов, не поддающихся непосредственному учету. Эти динамические свойства налагают определенные ограничения на передачу взаимодействия от одного процесса к другому, на периодичность колебаний и крутизну фронтов роста и спада.

**Варианты анализа.** Прямое прогнозирование курса доллара мало что может дать, сверх моделирования общей возрастающей тенденции. Предсказание резких изменений типа *черного вторника* (см. разд. 14.4) возможно только в непосредственной близости от такого рода событий и при наличии сильно развитой интуиции.

Если же попробовать прогнозировать фьючерсные котировки, то даже применение таких изопренных методов как модели Бокса–Дженкинса, вплоть до 12-го авторегрессионного порядка, дает нам лишь общее представление о динамике средней тенденции на очень короткое будущее, а различные конкретные генерации прогнозов существенно отличаются друг от друга. Возможно, что более надежные результаты дало бы моделирование и прогнозирование изменений курса (первая производная) или скорости этих изменений (вторая производная, характеризующая инерционность процессов).

Сначала попробуем построить интервальную автокорреляционную функцию временного ряда *fut*, выбрав в качестве сдвигаемого фрагмента 10 первых измерений, содержащих резкий выброс. Поэтому автокорреляционную функцию такого ряда можно рассматривать как переходную характеристику реакции системы на начальное возмущение.

### Результаты:

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ. Файл: spec.std  
Переменные: fut, fut Критич.значение=0.621  
\*E-1

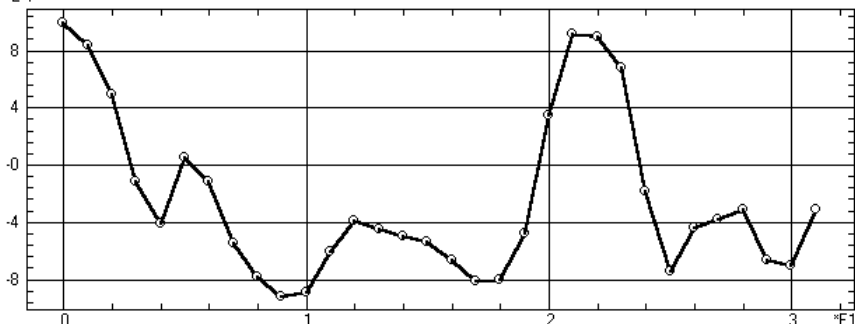


Рис. 9.3. Автокорреляционная функция фьючерсных котировок доллара. По горизонтальной оси — величина лага



**Обсуждение.** Как следует из рис. 9.3 автокорреляционная функция на первых 18 лагах достаточно монотонно уменьшается в область отрицательных значений, т. е. динамика временного ряда становится все более обратной по сравнению с начальным участком.

Однако на 18—20 лаге наступает резкий, но короткий переход к практически 100%—ной коррелированности с начальным участком. Такая задержка и характер отложенной реакции может служить важным показателем свойств исследуемой среды.

**Продолжение анализа.** Построим теперь интервальную кросскорреляционную функцию, отражающую влияние процесса *fut* на процесс *dol*.

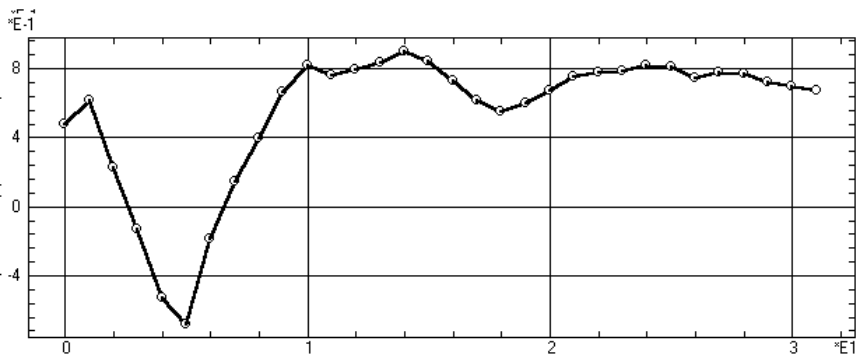


Рис. 9.4. Кросскорреляционная функция фьючерсных котировок доллара на курс доллара

**Обсуждение.** Полученная кросскорреляционная функция (рис. 9.4) достаточно монотонна, кроме начального отрицательного выброса, что подтверждает вполне очевидное заключение: фьючерсные котировки доллара мало влияют на сам курс доллара.

**Продолжение анализа.** Построим теперь интервальную кросскорреляционную функцию, отражающую влияние процесса *dol* на процесс *fut*. Для сравнения вычислим также и классическую корреляционную функцию, а для большей наглядности совместим их числовые выдачи и графики.

### Результаты:

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ. Файл: spec.std Переменные: dol, fut  
Сдвиг Интерв.корр. Класс.корр. Крит.значение  
Критич.значение=0.621

0	0.483	-0.0036	0.304
1	0.29	-0.185	0.308
2	0.296	-0.237	0.312
3	0.0362	-0.305	0.316
4	0.00264	-0.289	0.321

5	0.489	-0.159	0.325
6	-0.0451	-0.226	0.33
7	-0.547	-0.3	0.335
8	-0.572	-0.269	0.34
9	-0.527	-0.216	0.345
10	-0.487	-0.144	0.351
11	-0.212	-0.0355	0.357
12	0.000573	0.0329	0.363
13	-0.00285	0.0442	0.369
14	-0.079	0.0304	0.376
15	-0.39	-0.0492	0.383

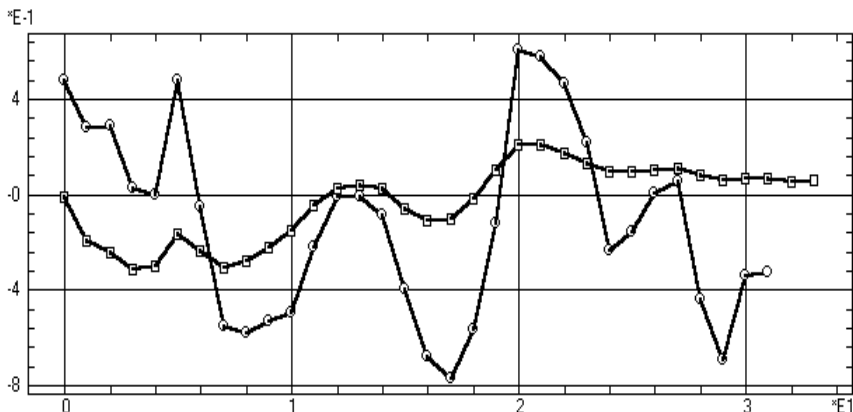


Рис. 9.5. Кросскорреляционная функция курса доллара на его фьючерсные котировки (квадраты — классическая корреляция, круги — интервальная корреляция)

**Обсуждение.** Интервальная корреляционная функция (рис. 9.5<sup>1</sup>) достоверно воспроизводит отложенную на 18–20 лагов резкую положительную реакцию фьючерсных котировок на начальный скачок курса доллара, отмеченную еще в автокорреляционной функции. Что же касается классической корреляционной функции, то по ее графику нельзя сделать никаких существенных выводов, поскольку все ее значения не отличны от нуля, а колебания незначительны.

Полученные статистические результаты дают нам несравненно более значимые материалы для выработки финансовой стратегии, чем моделирование и прогнозирование, предупреждая о том, что следует обращать особое внимание на динамику процессов спустя 18–20 дней после очередного возмущения. Эти же материалы дают пищу

<sup>1</sup> В качестве технической детали отметим, что совместный график корреляций получен следующим образом: вычислены две корреляционные функции с их индивидуальными графиками, данные с каждого графика перенесены в электронную таблицу нажатием инструментальной кнопки «СохрГраф» (в результате в электронной таблице появились две пары новых переменных), после чего построен функциональный график этих переменных.

специалистам для размышлений о физических свойствах среды, передающей взаимодействие процессов с такими большими задержками.

*Что дальше?* Здесь мы рассмотрели экстремальный случай реакции системы на резкое начальное возмущение. Дальнейший анализ может развиваться в направлении исследования взаимодействия процессов в обычных условиях, не содержащих резких колебаний. Оставляем эту задачу в качестве учебной читателям.

## 9.3. Спектральный анализ

### Пример 1

**З а д а ч а.** Необходимо проанализировать временную динамику авиаперевозок. Рассмотрим ежемесячные данные о расстояниях, пройденных самолетами Великобритании за шесть последовательных лет (переменная FLIGHT файл SPEC, в тысячах миль).

°E3

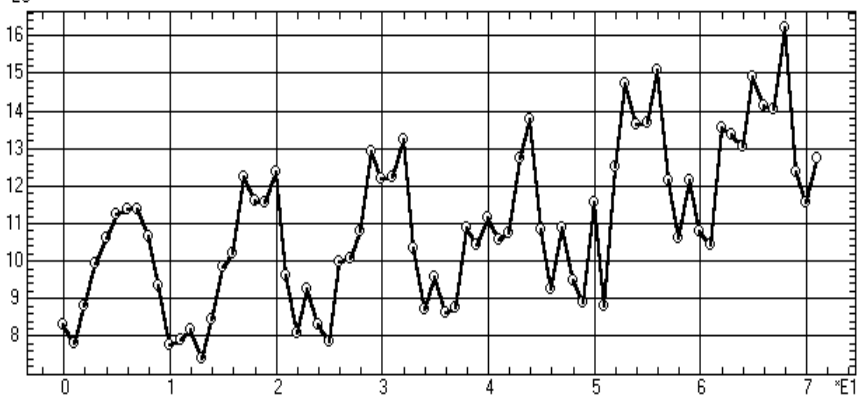


Рис. 9.8. График процесса *Flight* — ежемесячные данные о расстояниях, проходимых самолетами Великобритании

Сначала проведем визуальный анализ этого временного ряда (рис 9.8). Как можно заметить, эти данные характеризуются: а) линейным трендом с общей нестационарной тенденцией к возрастанию значений; б) сезонными колебаниями с периодом 12 мес. Для того чтобы нам исследовать различные типы периодичности в рассматриваемом процессе методом спектрального анализа, необходимо предварительно провести нормирование данных в *Блоке преобразований* (разд. 3.4) и обеспечить стационарность процесса удалением линейного тренда посредством выполнения процедуры простой регрессии с записью остатков в матрицу данных. Эти остатки и используются в дальнейшем анализе.

В результате этих операций мы получаем исходные данные для спектрального анализа, изображенные на графике рис. 9.9.

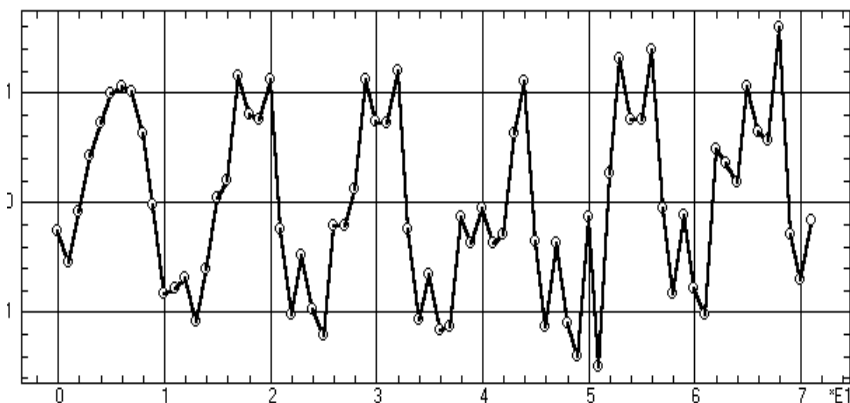


Рис. 9.9. График процесса *Flight* после удаления линейного тренда и нормализации

Проведем теперь спектральный анализ этого процесса.

### Результаты:

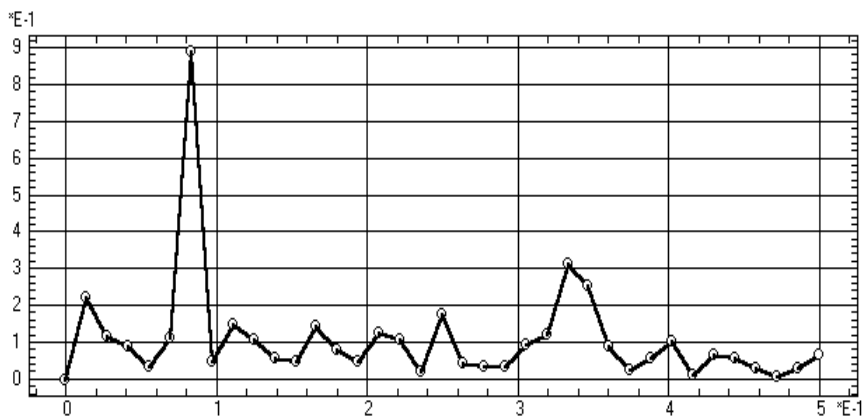
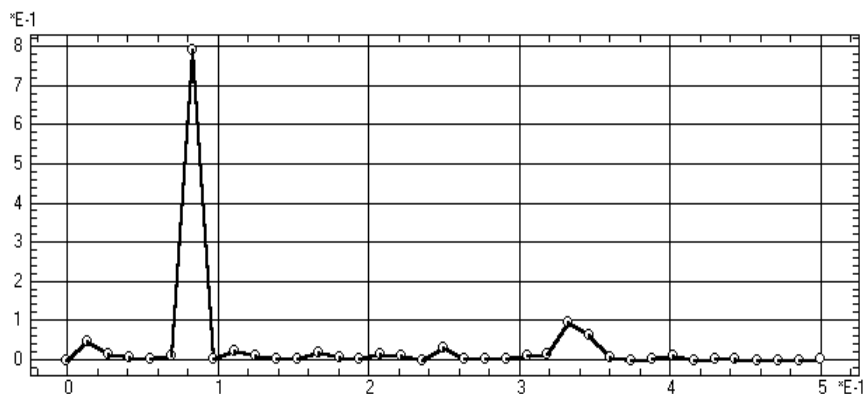
СПЕКТРАЛЬНЫЙ АНАЛИЗ. Файл: spec1.std Переменные: flight, flight

Частота	Период	Амплитуда	Фаза
0	0	0.00352	180
0.0139	72	0.221	-9.15
0.0278	36	0.119	82.1
0.0417	24	0.0886	-107
0.0556	18	0.0336	-70.9
0.0694	14.4	0.114	-113
0.0833	12	0.89	161
0.0972	10.3	0.0491	-36.4
0.111	9	0.148	-37.7
0.125	8	0.108	61.2
0.139	7.2	0.0581	-148

. . . . .

**Обсуждение:** В полученном спектре (рис. 9.10) виден доминирующий пик на частоте с периодом 1 год (что можно было бы наблюдать и на графике корреляционной функции). Однако вместе с этим наблюдается и второй, меньший пик для периода 0,3333 (4 месяца). Он достаточно удален от основного пика, поэтому маловероятно его появление вследствие эффекта вытекания мощности.

Для сравнения приведем периодограмму (рис 9.11), которая заменяет АЧХ в большинстве западных статистических пакетах (*StatGraphics*, *SPSS*, *Statistica*). Как можно легко заметить, этот график не дает никакой информации о любых спектральных составляющих, кроме подавляющего основного пика.

Рис. 9.10. Амплитудно–частотная характеристика процесса *Flight*Рис. 9.11. Периодограмма процесса *Flight*

*Продолжение анализа.* Если мы хотим провести более тщательное исследование процесса, не связанное с сезонностью, то необходимо повторить анализ с предварительным удалением частоты  $1/12$  сезонным фильтром  $1-B^{12}$  и частоты  $4/12$  одним из фильтров  $(1+B+B^2)/3$  или  $1-\sqrt{3}B+B^2$  с использованием средств разд. 9.3.

В результате этих операций график исходных данных приобретает вид, приведенный на рис 9.12. А после выполнения спектрального анализа мы получим график спектра, изображенный на рис. 9.13.

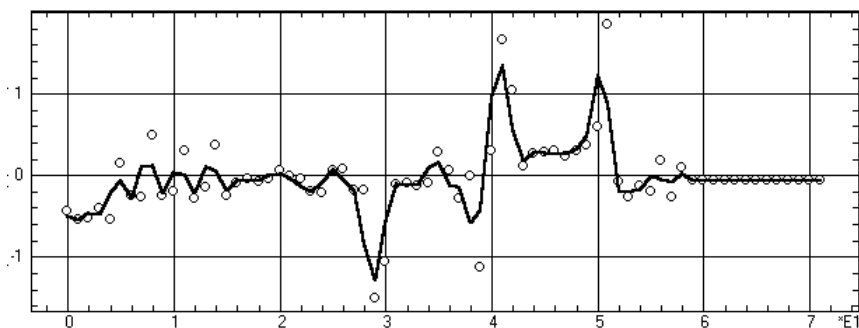


Рис. 9.12. График процесса *Flight* после фильтрации

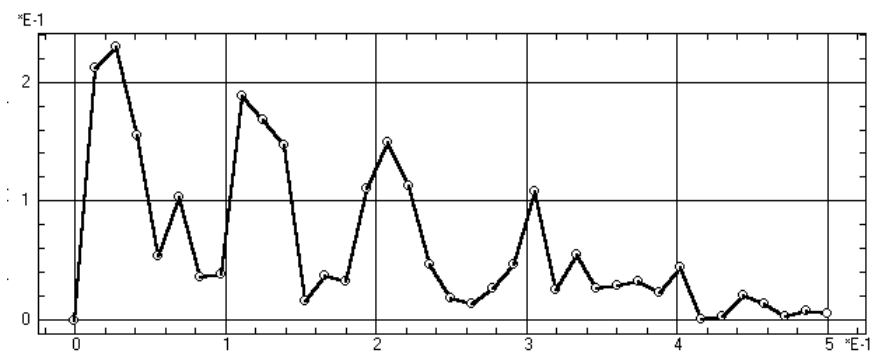


Рис. 9.13. Амплитудный спектр отфильтрованного процесса *Flight*

**Обсуждение:** Как можно заметить по графику спектра, в исследуемом процессе после проведенных фильтраций осталась, в основном, только случайная компонента и ее спектр сосредоточен преимущественно на низких частотах. Дальнейший интерес может представлять сравнение автокорреляционных функций, соответствующих вышерассмотренным спектрам.

Кроме того, для различных месяцев среднее значение и дисперсия могут существенно различаться. Поэтому можно проверить, что получится, если осуществить предварительную сезонную стандартизацию по месяцам. Продолжение анализа примера см. в следующем разделе.

## Пример 2

**Задача.** Необходимо исследовать временные взаимосвязи между числом отложенных яиц и числом взрослых насекомых, измеренных с периодичностью 2 дня (переменные *Eggs* и *Insects* в файле *SPEC*). Объединенный график этих двух временных рядов приведен на рис. 9.14.

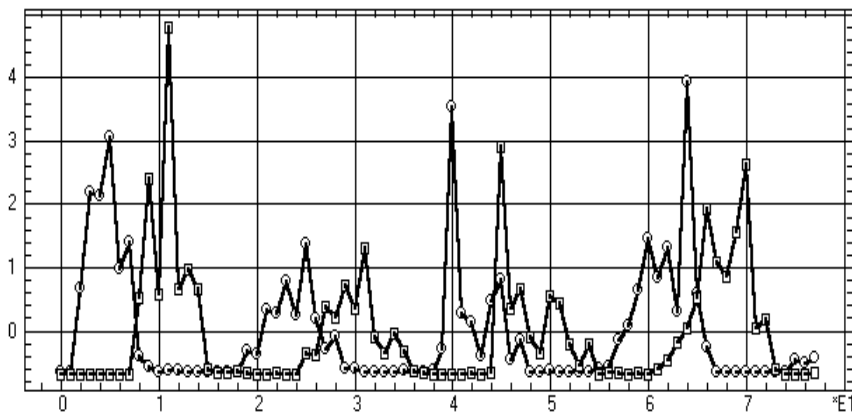


Рис. 9.14. График временных рядов *Eggs* (круги) и *Insects* (квадраты)

Как можно заметить, временные ряды не имеют заметного тренда и поэтому перед спектральным анализом можно провести только стандартизацию каждого ряда.

### Результаты:

СПЕКТРАЛЬНЫЙ АНАЛИЗ. Файл: spec1.std Переменные: eggs, insects

Частота	Период	Амплитуда	Фаза	Когерент.	Передат.функ.	
0	0	6.7E-5	0	1	0.00812	0.00826
0.0128	78	0.0211	29.6	1	0.156	0.136
0.0256	39	0.00339	67.1	1	0.0475	0.0712
0.0385	26	0.141	83	1	0.344	0.41
0.0513	19.5	0.822	110	1	0.919	0.894
0.0641	15.6	0.017	136	1	0.109	0.156
0.0769	13	0.00738	-173	1	0.0664	0.111
0.0897	11.1	0.131	-174	1	0.342	0.383
0.103	9.75	0.0915	-140	1	0.346	0.264

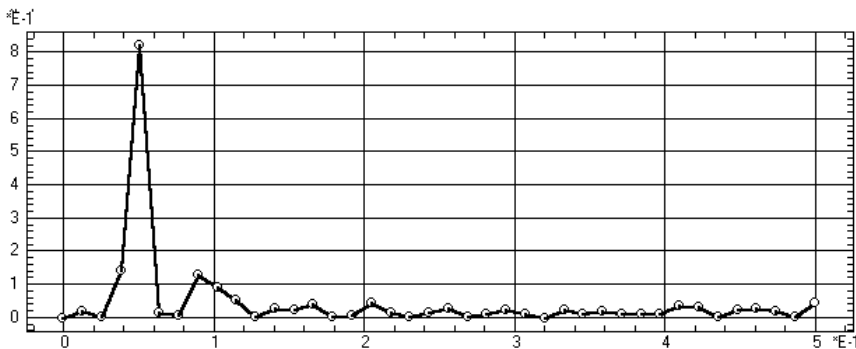


Рис. 9.15. Кроссспектр временных рядов *Eggs* и *Insects*

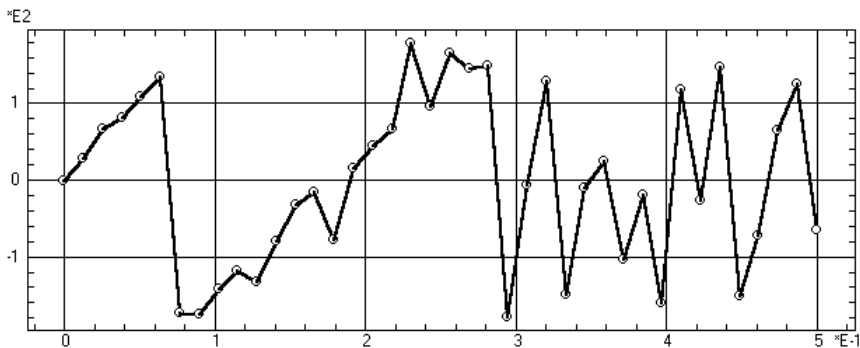


Рис. 9.16. График фазы кросс-спектра для рядов *Eggs* и *Insects*

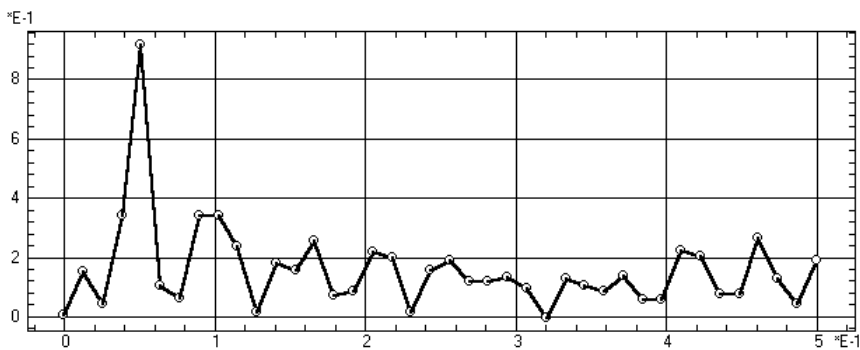


Рис. 9.17. График передаточной функции для рядов *Eggs* и *Insects*

**Обсуждение:** Как видно из рис. 9.15, мощность кросс-спектра сосредоточена преимущественно на низких частотах (при дополнительном анализе с усреднением легко убедиться, что на этих частотах и значения когерентности являются устойчиво высокими, что позволяет принять с доверием оценки фазы и передаточной функции).

Выделяется пик с периодом 39 дней, что можно интерпретировать как среднее время развития и жизни насекомого.

Фаза (рис. 9.16) имеет явный линейный тренд на низких частотах, что иллюстрирует систематическое смещение рядов друг относительно друга (от момента снесения яйца до превращения во взрослое насекомое). В то же время передаточная функция является относительно постоянной в этой области. Поскольку рассматриваемые процессы связаны обратной связью, интерпретация передаточной функции должна выполняться с осторожностью.



Эти результаты позволяют выдвинуть гипотезу о том, что на число взрослых насекомых в каждом следующем поколении влияет только число яиц в предыдущем поколении, а не их распределение во времени.

Для более отчетливого выделения информации о взаимных связях двух процессов можно провести повторный анализ с различными степенями сглаживания исходных процессов и результатов. Полезной при анализе может быть и кросс-корреляционная функция, поскольку наклон фазы в известной степени соответствует лагу (в данном случае равное  $-6$ ) максимума корреляционной функции. Поэтому можно вычислить спектр еще раз, выполняя выравнивание с параметром, близким к  $-6$ .

## 9.4. Сглаживание и фильтрация

### Пример 1

**Задача.** В разрушающем техническом эксперименте регистрировалось ускорение головы манекена водителя мотоцикла после столкновения его с неподвижным жестким препятствием (файл МТО). Необходимо исследовать эту зависимость.

**Обсуждение:** Обратимся сначала к визуальному анализу (рис. 9.20). Прежде всего бросается в глаза большая зашумленность экспериментальных данных, вызванная погрешностями измерений ускорения. Требуется некоторым образом аппроксимировать эту зависимость, чтобы высветить ее скрытую шумом природу и получить более точные оценки закона изменения ускорения.

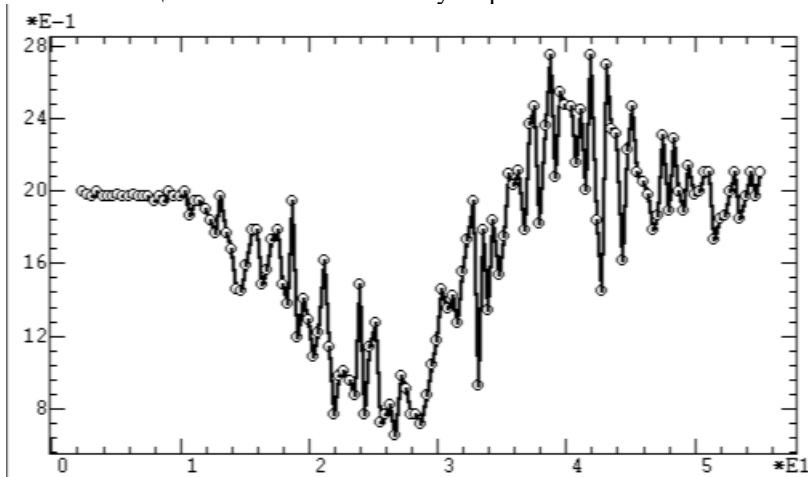


Рис. 9.20. Ускорение [g] головы водителя мотоцикла в зависимости от времени после столкновения с препятствием [ms]

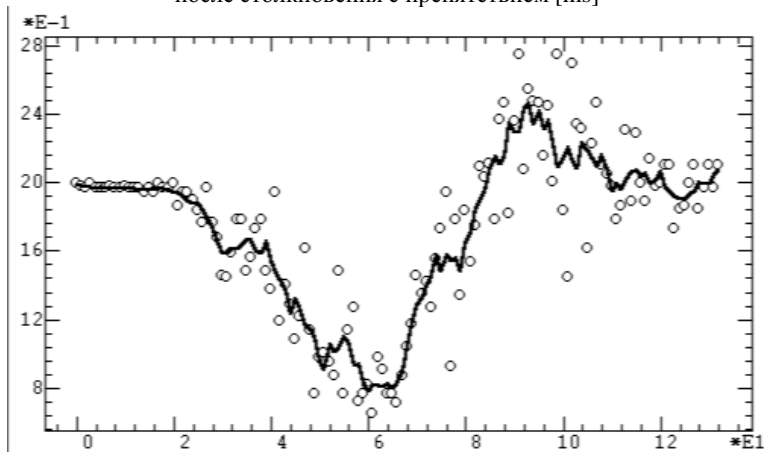


Рис. 9.21. Сглаживание скользящим средним по пяти точкам

Для столь зашумленных рядов более качественное сглаживание дают не методы скользящего среднего, а методы робастного сглаживания, менее чувствительного в резким колебаниям процесса.

Действительно, как легко видеть по сравнению графиков результатов (рис. 9.21, 9.22) метод робастного сглаживания Хубера ( $W=0,05$ , ширина=5, константа=10, треугольное окно) дает намного более чем сглаживание скользящим средним по пяти точкам.

Еще более впечатляющий эффект (рис 9.23) дает применение сглаживающего сплайна с числом шагов 2 и коэффициентом 0,05 (см. разд. 4.2).

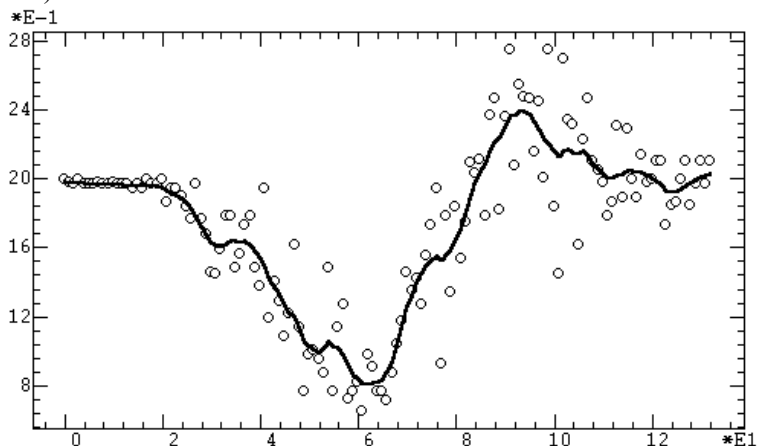


Рис. 9.22. Робастное сглаживание

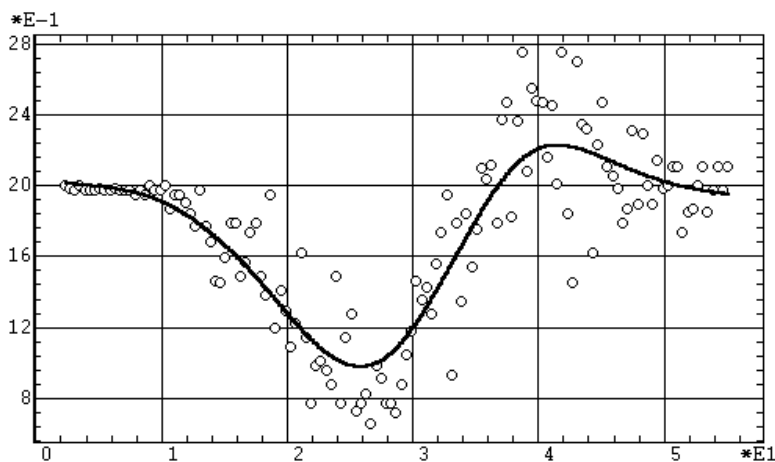


Рис. 9.23. Сглаживание сплайном

## Пример 2

**Задача.** На примере временной динамики авиаперевозок (рис. 9.8) следует визуально оценить результаты различных методов сглаживания и фильтрации.

Продолжение анализа данного примера см. в следующем разделе.

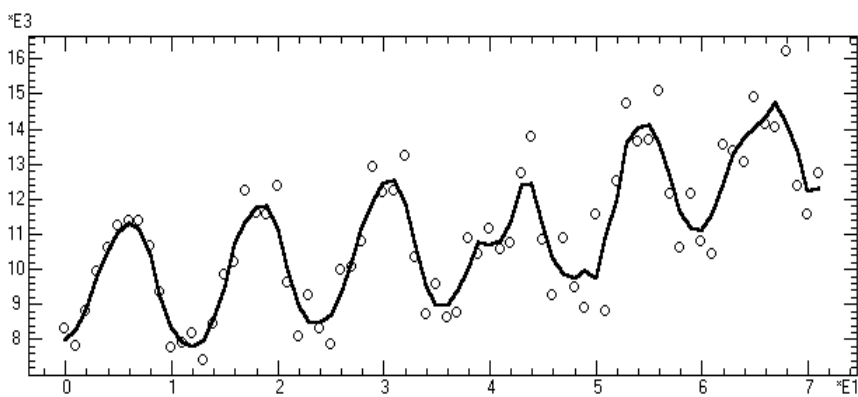


Рис. 9.24. Сглаживание ряда *Flight* скользящим средним по трем точкам (кружки – экспериментальные точки)

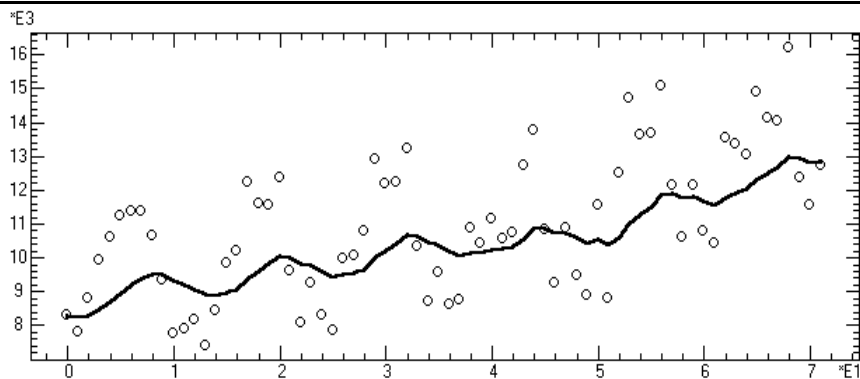


Рис. 9.25. Экспоненциальное сглаживание ряда *Flight* с коэффициентом  $w=0,1$ , способствующим выделению тренда

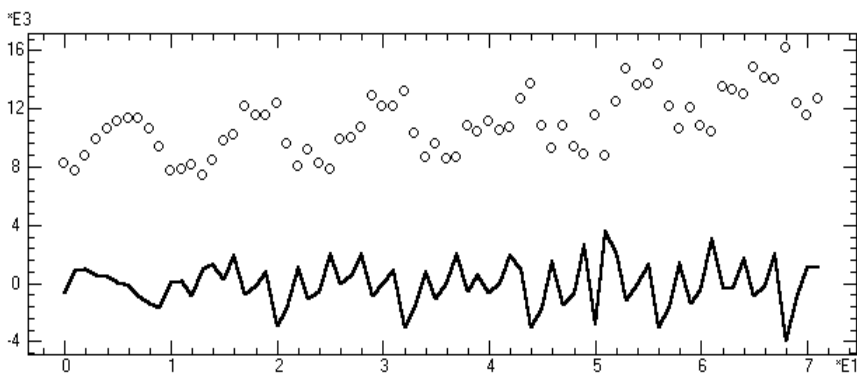


Рис. 9.26. Фильтрация ряда *Flight* дифференцированием

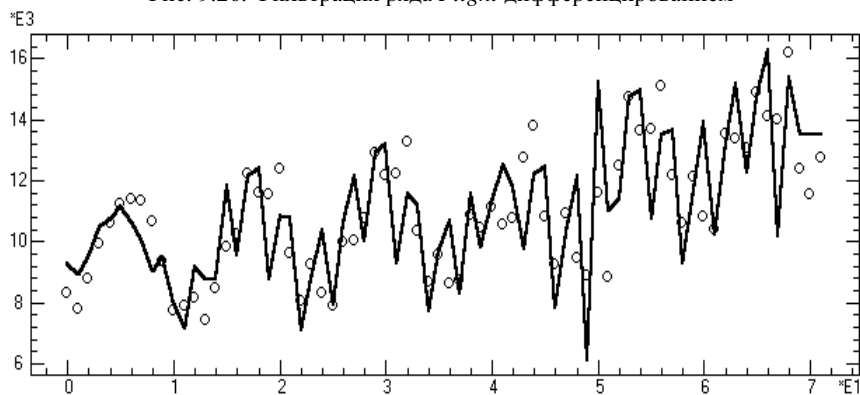


Рис. 9.27. Фильтрация дифференцированием–интегрированием

## 9.5. Авторегрессионные модели

### Пример

**З а д а ч а.** Построим ARIMA-модели для временного ряда авиаперевозок (см. рис. 9.8) после его нормирования и удаления линейного тренда (см. рис. 9.9).

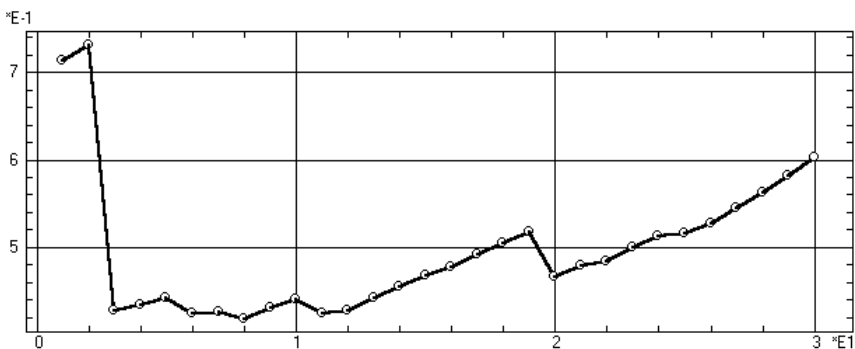


Рис. 9.29. График ошибки прогноза AR-модели от числа параметров

**Обсуждение:** Рассмотрим первый результат анализа — график ошибки прогноза на рис. 9.29. Как можно заметить, минимум ошибки прогноза имеет место при 3–4 параметрах в модели авторегрессии. Поэтому в качестве начальных условий можно указать четыре параметра авторегрессионной модели и 0 параметров в модели скользящего среднего, поскольку в анализируемом ряде шумовая составляющая очевидно невелика по сравнению с закономерной периодической составляющей и, следовательно, в первом приближении можно ограничиться простейшей моделью белого шума.

### Результаты:

АВТОРЕГРЕССИЯ БОКСА-ДЖЕНКИНСА. Файл: spec1.std Переменная flight  
 Модель: AR=4, MA=0, дифф=0, хи2=5.12, ст.своб=7, значим=0.645  
 Гипотеза 0: <Модель адекватна временному ряду>

Коэффициенты авторегрессионной модели						
	a0	a1	a2	a3	a4	a5
	-1.6E-5	0.662	-0.196	0.527	-0.74	
Коэффициенты модели скользящего среднего						
Шум	b1	b2	b3	b4	b5	b6
	0.21					
Время	Ср.прогн	Ст.ошиб	Довер.инт	Прогн1	Прогн2	Прогн3
1986	-1.305	0.5239	1.026	-1.887	1.033	-1.919
1986	-0.9899	0.6284	1.231	-1.239	0.8491	-0.9013
1986	0.03444	0.6411	1.255	-0.4842	0.256	-0.528

1986	-0.3534	0.7046	1.38	-1.765	-0.09418	-0.6128
1986	0.203	0.7055	1.382	-0.1943	0.5309	2.414
1986	0.954	0.7557	1.48	0.9131	0.5516	2.497
1986	0.3801	0.7639	1.496	0.3331	0.3516	1.029
1987	0.4332	0.8058	1.578	0.9709	0.4742	2.571
1987	0.565	0.8479	1.66	-0.07651	0.164	0.9153
1987	-0.2161	0.8481	1.661	-0.9155	0.6421	-0.8492
1987	-0.3066	0.8481	1.661	-0.4995	0.6371	0.215
1987	-0.1833	0.8499	1.664	-0.8315	0.7391	-0.9581

Корреляция остатков Дурбина-Ватсона=1.11, значимость=0.344  
Гипотеза 0: <Модель адекватна временному ряду>

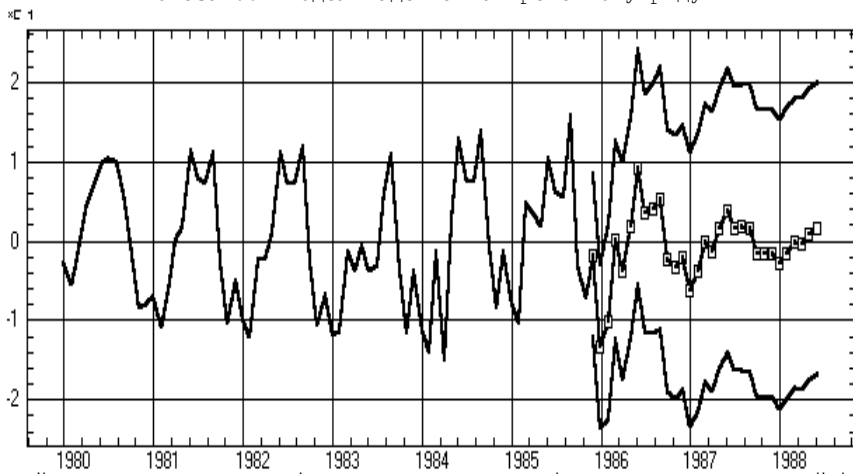


Рис. 9.30. График временного ряда со средним прогнозом и 95%-ным доверительным интервалом ARIMA-прогноза

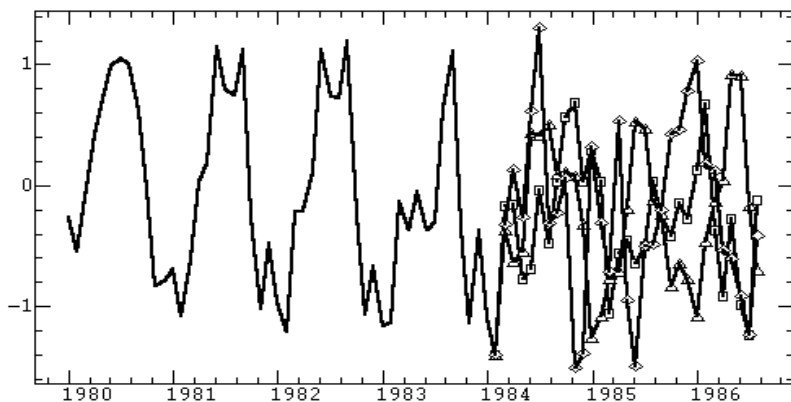


Рис. 9.31. Временной ряд с тремя генерациями ARIMA-прогноза

**Обсуждение:** Приведенная выдача результатов показывает, что произведенный выбор параметров позволил получить достаточно адекватное

описание временного ряда четырехпараметрической моделью, с помощью которой произведен прогноз изменения интенсивности авиаперевозок на 30 месяцев вперед. Далее выдаются графики прогноза и остатков.

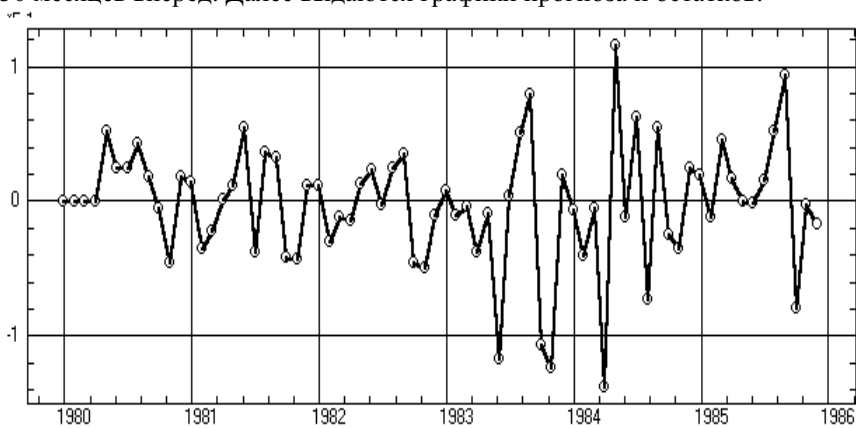


Рис. 9.32. График распределения ARIMA-остатков

Как можно заметить из рис. 9.30, амплитуда сезонного изменения среднего прогноза существенно уменьшается с течением времени (в соотнесении с более стабильной сезонной вариабельностью исходного ряда), а зона доверительного интервала (т. е. ошибки прогноза) достаточно широка, что говорит о недостаточной точности и устойчивости модели. Значительное влияние шумовой составляющей выявляется и на рис 9.31, где три генерации прогнозов существенно различаются между собой вплоть до прямо противоположных предсказаний в некоторых точках. Величина остатков (рис. 9.32) также заметно возрастает от начала к концу ряда. Все выявленные недостатки в большой степени относятся собственно к рассматриваемому методу.

В качестве варианта дальнейшего анализа можно предложить предварительную трансформацию временного ряда методом логарифмирования, что позволяет сместить интерес исследования с абсолютного изменения числа пассажиров на его относительное изменение.

Для достижения стационарности можно также попробовать применение дифференциального фильтра для удаления линейной тенденции и сезонного 12-месячного фильтра для устранения сезонных изменений.

Далее можно построить модели для различных отрезков временного ряда с целью выяснения однородности данных в смысле применимости одной и той же модели к первой и второй их половине.

Продолжение анализа данного примера см. в следующем разделе.

## 9.6. Фурье–модели

### Пример 1

**Задача.** Данный пример предназначен для сравнения фурье–моделей с другими методами анализа временных рядов. Поэтому в качестве исходных данных используем хорошо знакомый в данной главе временной ряд изменения авиаперевозок (см. рис. 9.9). Внимательное изучение исходного временного ряда показывает, что в нем с течением времени нарастают амплитуды высокочастотных нерегулярных составляющих, существенно искажающие главный годичный ритм (нестационарность).

**Шаг 1.** Поскольку временной ряд уже нормирован и центрирован предварительная коррекция модели (тренд, скачки) не нужна. Амплитудный спектр процесса приведен на рис. 9.10. Для начала произведем только фильтрацию высокочастотных составляющих спектра в диапазоне от 20 до 36. Для этого в бланке параметров фурье–модели (рис. 9.34) следует заполнить позиции верхнего и нижнего фильтров, остальные позиции оставим нулевыми.

### Результаты:

ФУРЬЕ–МОДЕЛИ Файл: spec1.std  
Ст.откл.шума=0,343

Переменная flight

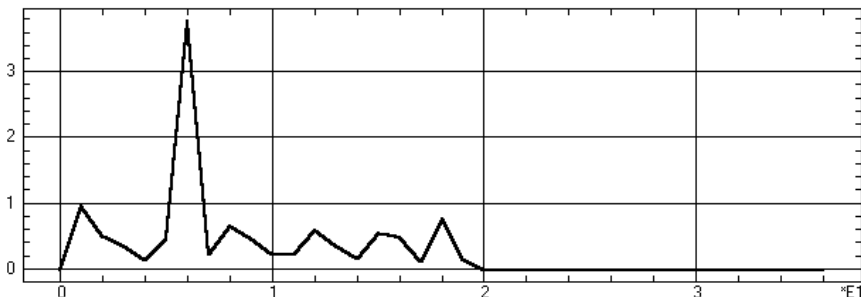


Рис. 9.35. График спектральной модели после удаления высокочастотных составляющих



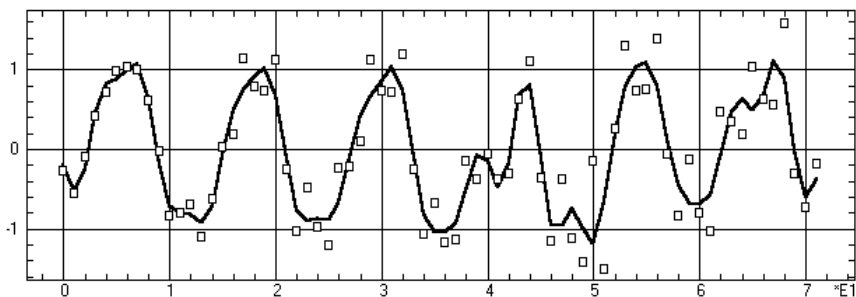


Рис. 9.36. График фильтрации временного ряда авиаперевозок

**Обсуждение:** В результате фильтрации спектр процесса приобрел вид рис. 9.35. Как можно видеть по графику результата фильтрации (рис. 9.36), удаление высокочастотных составляющих из спектра действительно привело к желаемому сглаживанию быстрых колебаний временного ряда. Полученный результат близок к результату фильтрации ряда скользящим средним (см. рис. 9.24). Удаленные компоненты соответствуют шуму (характеризует различие между моделью и временным рядом) со стандартным отклонением 0,343.

**Шаг 2.** Удалим теперь из спектра основную спектральную гармонику (сезонный компонент) в диапазоне частот 4–8.

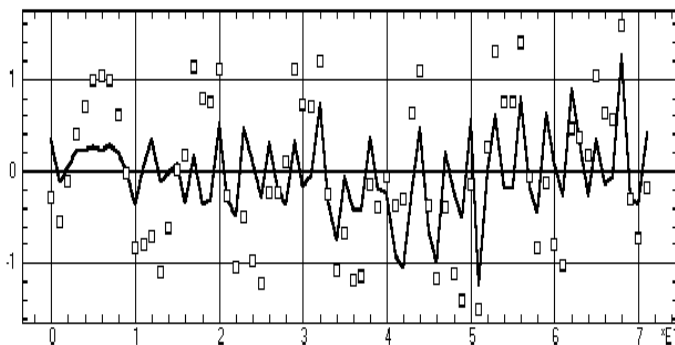


Рис. 9.37. График спектральной модели после сезонной фильтрации

**Обсуждение:** Сравним полученный результат (рис 9.37) с результатом сезонной фильтрации из разд. 9.4 (см. рис. 9.27). Как можно заметить, фурье-фильтрация обеспечивает удаление сезонной составляющей без резких выбросов, характерных для рис. 9.27, кроме того результат охватывает всю длину временного ряда, без присущего сезонному дифференцированию «обрезания» конца ряда.

**Шаг 3.** Используем теперь нефильтрованную («чистую») фурье–модель для прогнозирования авиаперевозок на 30 шагов, т. е. на два с половиной года. Для этого в бланке рис. 9.34 заполним только поле прогноза.

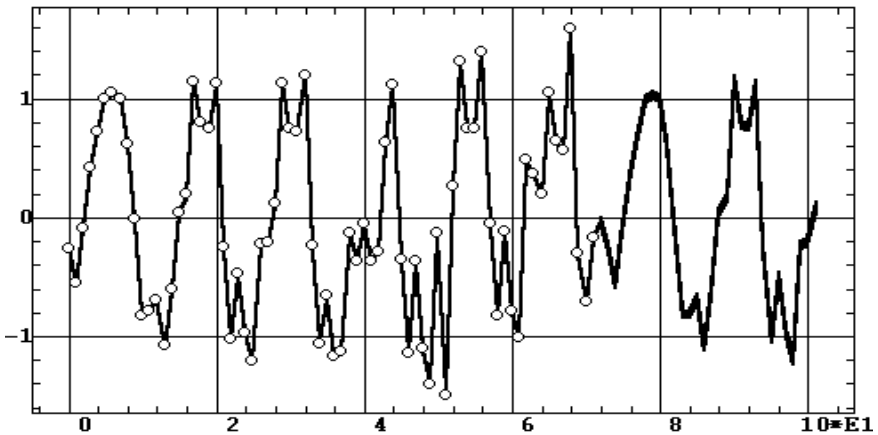


Рис. 9.38. Временной ряд авиаперевозок с прогнозом на 30 шагов

**Обсуждение:** Как видно из полученного результата (рис 9.38), чистая фурье–модель полностью повторяет временной ряд и ее прогноз является повторением начального участка процесса.

**Шаг 4.** Изменим модель, удалив из спектра малоамплитудные составляющие, не превышающие 15% от максимума.

**Результаты** (сокращенно):

```

Файл: spec1.std      Переменная flight
      Параметры Фурье-модели
      Амплитуда      Период      Фаза
      0.00571        0          180
      0.217          73         -9.34
      0.894         12.2        161
      0.145          9.13       -38.8
      0.145          6.08       -105
      0.172          4.06       -15
      0.313          3.04       71.3
      0.258          2.92      -152
  
```

Ст.отклонение остатков (различие модели и процесса)=0.0817

X	Yэксп	Yмодл	остаток	Ст.остат	Ст.ошиб	Довер.инт
0	-0.256	-0.457	0.202	0.706	0.296	0.582
1	-0.537	-0.212	-0.325	-1.12	0.295	0.581
2	-0.081	-0.166	0.0851	0.301	0.295	0.58
3	0.432	0.279	0.153	0.538	0.295	0.58
4	0.737	0.658	0.0789	0.279	0.294	0.579
5	1.01	0.659	0.347	1.21	0.294	0.579
6	1.06	0.963	0.097	0.342	0.294	0.578
7	1.02	1.25	-0.234	-0.807	0.294	0.578
8	0.635	0.786	-0.151	-0.519	0.293	0.577
9	-0.00816	-0.0947	0.0865	0.306	0.293	0.577
10	-0.822	-0.511	-0.311	-1.08	0.293	0.576

```

Хпрогн      Упрогн      Ст.ошиб      Довер.инт
72          -0.457      0.296        0.582
  
```

---

73	-0.212	0.296	0.583
74	-0.166	0.297	0.584
75	0.279	0.297	0.584
76	0.658	0.297	0.585
77	0.659	0.298	0.586
78	0.963	0.298	0.587
79	1.25	0.298	0.587
80	0.786	0.299	0.588

. . .

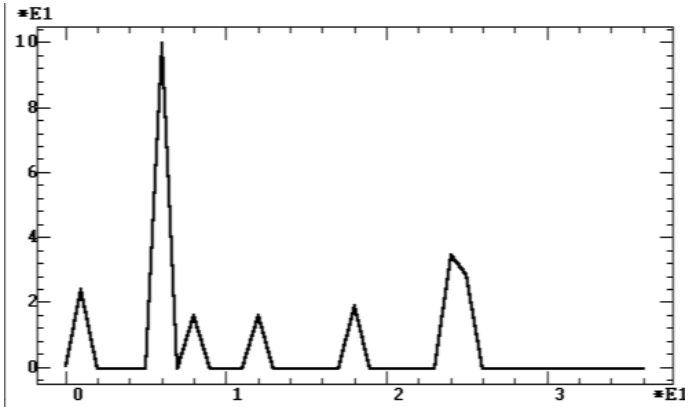


Рис. 9.39. График спектральной модели после удаления малоамплитудных составляющих

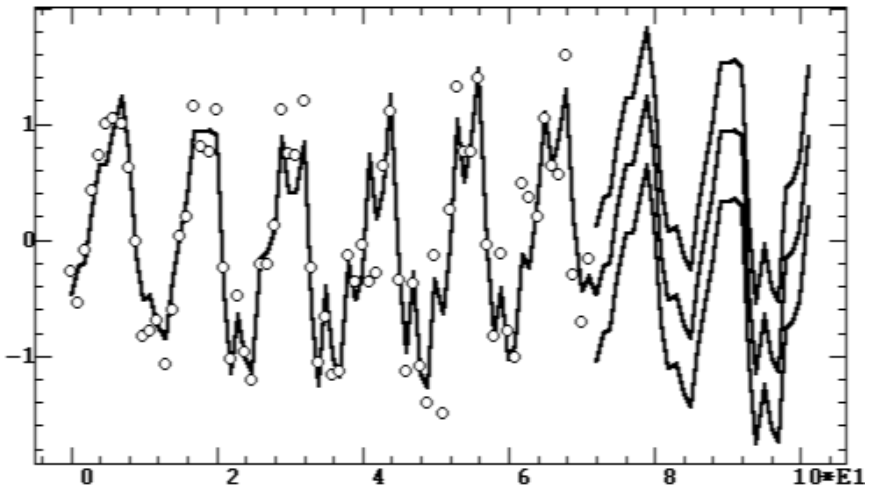


Рис. 9.40. Временной ряд авиаперевозок с фурье-моделью и с прогнозом на 30 шагов

*Обсуждение:* Здесь с иллюстративными целями приведена также и числовая выдача результатов с параметрами модели, анализом остатков и прогнозом. Амплитудный спектр модели приведен на рис. 9.39.

Как видно из результата моделирования и прогноза (рис 9.40), модель уже не повторяет временной ряд. По сравнению с рис. 9.38 появились высокочастотные колебания на начальном участке процесса, но снизилась их амплитуда на конечном участке. Однако прогноз остается повторением начального участка смоделированного ряда.

*Шаг 5.* Добавим к данной модели еще и адаптацию.

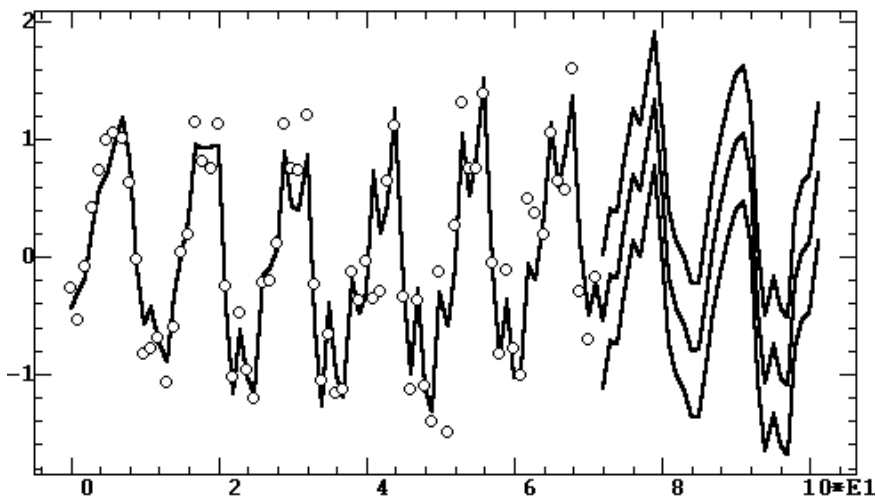


Рис. 9.41. Фурье-модель с адаптацией

*Обсуждение:* Как видно из рис 9.41 за счет адаптации само моделирование процесса по сравнению с рис. 9.40 более близко к чистой фурье-модели (см. рис. 9.38) и к исходному временному ряду, к тому же прогноз уже не повторяет начало временного ряда.

*Шаг 6.* Добавим теперь в модель еще и два усреднения.

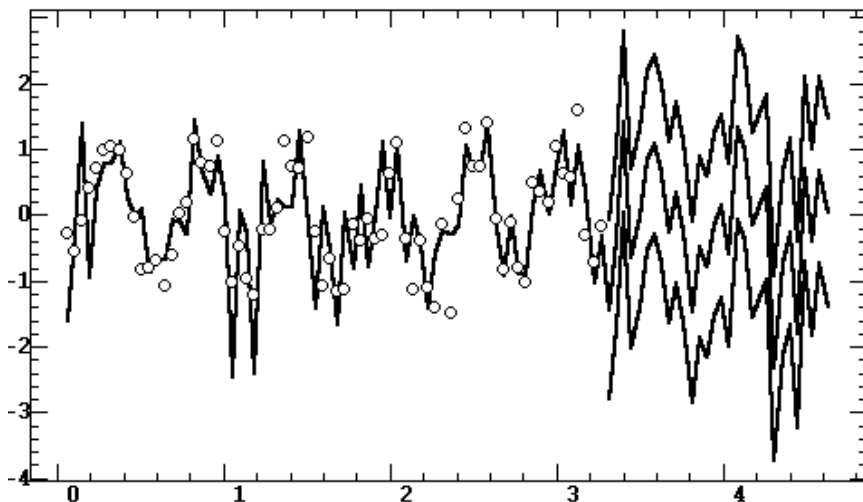


Рис. 9.42. Фурье–модель с адаптацией и двумя усреднениями

**Обсуждение:** Как видно из рис. 9.42 за счет усреднения по сравнению с рис. 9.41 в зоне прогнозирования более представлены высокочастотные колебания, характерные для последнего участка временного ряда. Поэтому в отношении прогноза эту модель следует признать более адекватной, хотя при моделировании начального участка авиаперевозок там появились не характерные для него высокочастотные колебания.

## Сравнение методов

**З а д а н и е:** Проведем сравнение предложенного фурье–метода с авторегрессионным подходом (разд. 9.5). Для этого исключим из рассматриваемого временного ряда его последнюю треть и сравним точность прогнозирования на этом участке, достигаемую с использованием каждого метода (рис. 9.38). Для фурье–модели используем уже опробованную схему двукратного усреднения с адаптацией и очисткой спектра.

### Результаты:

X	Y	Y <sub>фурье</sub>	Y <sub>arima</sub>	X	Y	Y <sub>фурье</sub>	Y <sub>arima</sub>
50	-0.1187	-0.861	-0.151	61	-1.003	-1.89	-0.661
51	-1.482	0.281	-0.0169	62	0.4955	-0.778	-0.27
52	0.2741	0.383	-0.0822	63	0.3761	0.543	0.0779
53	1.321	-0.0206	0.824	64	0.2025	0.623	0.0485
54	0.7636	0.453	0.769	65	1.06	0.141	0.272
55	0.7658	1.33	0.199	66	0.6503	0.375	0.403
56	1.405	1.16	0.373	67	0.5779	0.817	0.139
57	-4.042E-2	0.176	0.047	68	1.603	0.484	0.0277
58	-0.8149	-0.457	-0.62	69	-0.2873	-0.0613	-0.0376
59	-0.1097	-0.869	-0.505	70	-0.7006	-0.0925	-0.311

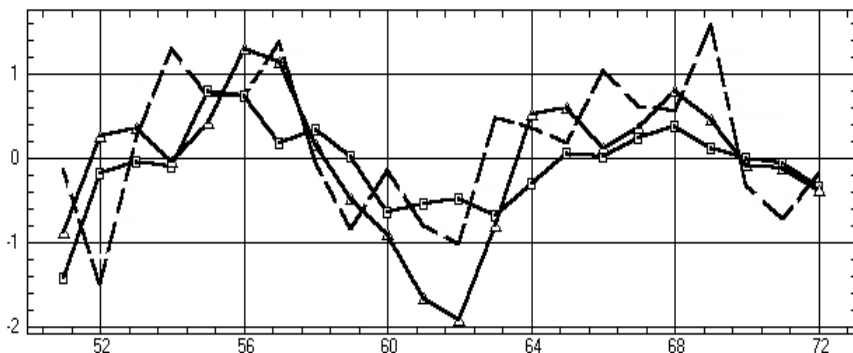


Рис. 9.43. Сравнение фурье-прогноза (треугольники) и ARIMA-прогноза (квадраты) с временным рядом авиаперевозок (пунктирная линия)

**Обсуждение:** В числовых результатах выше приведены истинные значения прогнозируемого временного ряда  $Y$  и прогностические значения фурье- и ARIMA-моделей;  $Y_{\text{фурье}}$ ,  $Y_{\text{arima}}$ . Для количественного сравнения обоих методов вычислим среднее стандартное отклонение  $S$  для разностей между прогностическими и истинными значениями ( $Y - Y_{\text{фурье}}$ ,  $Y - Y_{\text{arima}}$ ) и получим:  $S = 0,578$  для фурье-модели и  $S = 0,641$  для ARIMA-модели.

Таким образом фурье-модель дает прогноз в среднем на  $9,8\% = (0,641 - 0,578) / 0,641 \cdot 100$  точнее, чем широко популярная, но алгоритмически неизмеримо более сложная и трудная для понимания ARIMA-модель.

**Заключение:** На основании выводов по рассмотренным примерам можно сделать следующий вывод: предложенные фурье-модели являются многоцелевым инструментом исследования и в каждой из своих областей применения дают результаты не хуже, а зачастую – и лучше, чем популярные и давно известные аналоги.

## Пример 2

**Задача.** Попробуем смоделировать и спрогнозировать натуральный (непреобразованный) временной ряд изменения урожайности зерновых в СССР из примера 1 к разд. 10.3. Поскольку, как там показано, для этого ряда более подходит параболическая модель тренда по сравнению с линейной, то для сравнения используем обе эти модели. Кроме того в этом процессе не наблюдается явных скачков, что следует из самой природы процесса (примеры с удалением скачков рассмотрены разд. 14.4).

**Результаты** (сокращенно и без числовой выдачи):

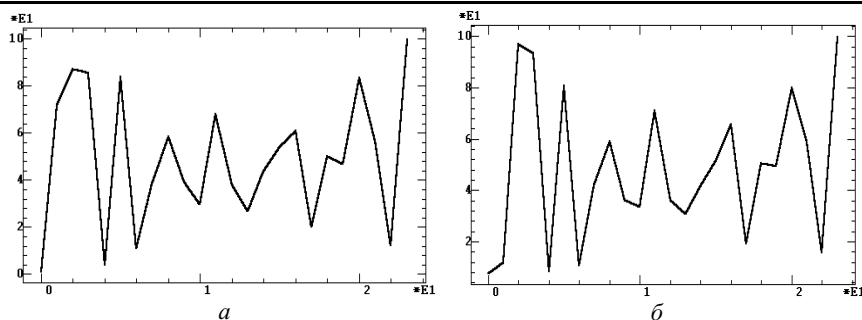


Рис. 9.44. Спектры изменения урожайности зерновых после удаления тренда:  
а — линейный тренд; б — полиномиальный тренд

**Обсуждение:** Как видно из графиков полных спектральных моделей (рис. 9.44) малозначимые компоненты можно отсечь на уровне 30% от амплитуды максимальной гармоники. В связи с небольшой длиной временного ряда выполним только два усреднения с адаптацией, дополнительной фильтрации делать не будем, выполним прогноз изменения урожайности зерновых на 20 лет вперед, т. е. до 2009 г.

**Результаты** (сокращенно и без числовой выдачи):

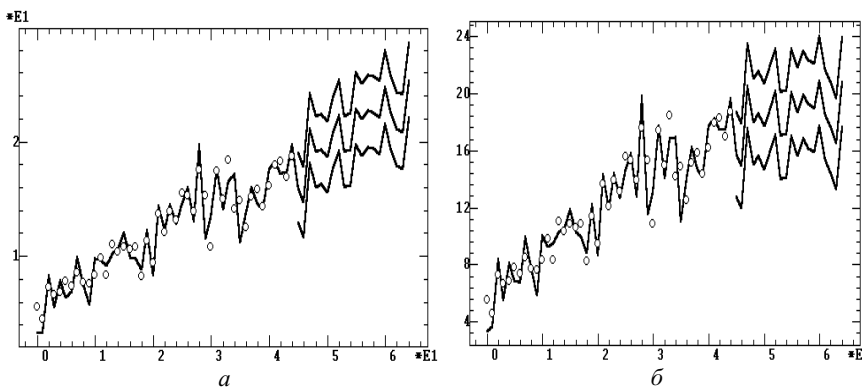


Рис. 9.45. Модель и прогноз урожайности зерновых в СССР:  
а — линейный тренд; б — полиномиальный тренд

**Обсуждение:** Как видно из графиков моделей и прогнозов, параболическая модель (рис. 9.45) предсказывает в среднем практически полную стагнацию урожайности на 20 лет. Линейная же модель (рис. 9.45, а) предсказывает в среднем рост приблизительно линейного характера, на котором, однако, внимательным взглядом можно заметить продолжение замедления роста, характерное для последней трети процесса. Тем самым линейную модель тренда в данном случае следует признать более адекватной средним тенденциям временного ряда. Почему это

происходит, несмотря на лучшее соответствие временного ряда параболической модели, выявленное в примере 1 разд. 10.3. Как можно заметить из сравнения спектров рис. 9.44 в линейной модели намного более выражена самая низкочастотная гармоника. И именно она компенсирует сравнительно меньшую адекватность модели линейного тренда, добавляя к нему длинно периодическую синусоиду. Это было уже отмечено в основной части данного раздела.

Еще следует отметить, что обе модели проявляют существенные высокочастотные колебания урожайности, характерные для последней половины временного ряда, определенную двумя заказанными в модели усреднениями.

### Пример 3

**Задача.** Рассмотрим ежемесячную динамику цен на нефть Брент с 2001 по 2004 гг.<sup>1</sup> (рис. 9.46, файл OIL). Этот процесс визуально также имеет явную нелинейность, но в отличие от предыдущего примера не затухающую, а возрастающую. Поэтому здесь также попробуем сравнить две модели линейную и параболическую. Кроме того, в этом процессе не наблюдается явных скачков (примеры с удалением скачков рассмотрены разд. 14.4).

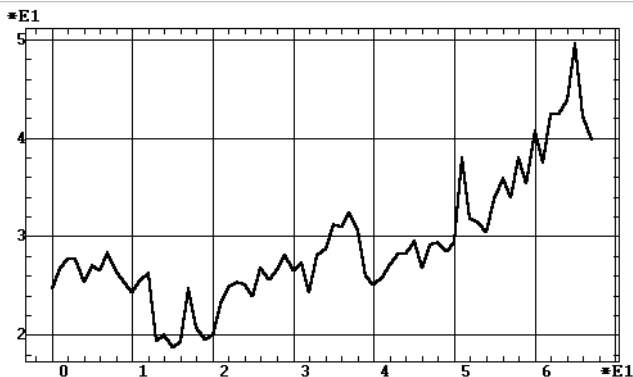


Рис. 9.46. Ежемесячная динамика цен на нефть Брент с 2001 по 2004 гг.

**Обсуждение:** Как видно из графиков полных спектральных моделей (рис. 9.47) малозначимые компоненты можно попробовать отсечь на уровне 10%. При линейном тренде преобладающую амплитуду имеет низкочастотная составляющая вследствие высокоамплитудных и длинно периодических колебания остатков относительно тренда. Полиномиальная модель намного более адекватна процессу и в ней в

<sup>1</sup> Данные предоставлены научным сотрудником ВНИИ внешнеэкономических связей В.А. Ярных



равной степени представлены низко- и высокочастотные составляющие. Остальные параметры спектральной модели оставим аналогичными примеру 2 и выполним прогноз на два последующих года.

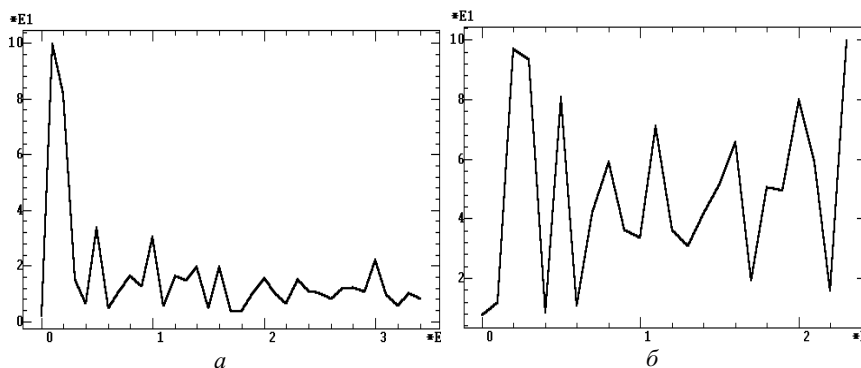


Рис. 9.47. Спектры цен на нефть Брент после удаления тренда:  
 $a$  — линейный тренд;  $b$  — полиномиальный тренд

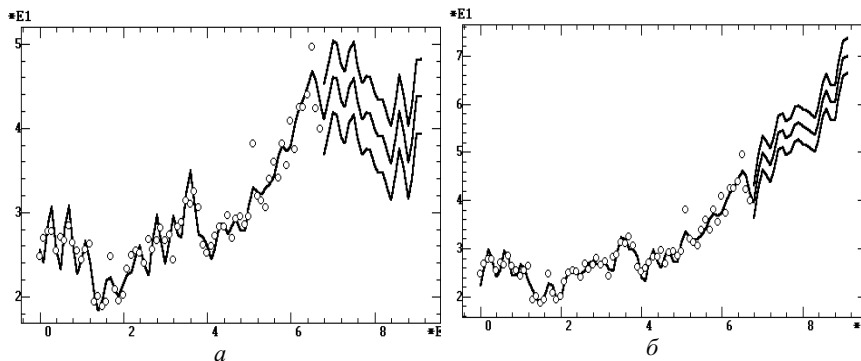


Рис. 9.48. Модели и прогнозы цен на нефть Брент:  
 $a$  — линейный тренд;  $b$  — полиномиальный тренд

**Обсуждение:** Как видно из графиков моделей и прогнозов, линейная модель (рис. 9.48,  $a$ ) за счет сильной низкочастотной составляющей в спектре (рис. 9.47,  $a$ ) предсказывает в снижении цен на нефть в последующие два года. В противоположность этому параболическая модель (рис. 9.48,  $b$ ) предсказывает продолжение тенденции примерно линейного роста цен на нефть. Однако из содержательных соображений в начале 2005 г. (т. е. не имея апостериорной информации) можно было бы сказать, что оба этих прогноза недостаточно реалистичны, и истина должна лежать где-то посередине. Возможно, в качестве модели тренда данного процесса более подошла бы парабола третьей или более высокой степени.

# РЕГРЕССИОННЫЙ АНАЛИЗ

## 10.2. Сравнение двух линий регрессии

### Пример

**З а д а ч а.** При ревматоидном артрите из-за болезненности суставов происходит частичная атрофия мышц, что сказывается на физической силе. Чтобы проверить статистическую значимость этих изменений, в группе больных и в группе здоровых людей измерялись зависимости силы сжатия кисти от поперечного сечения предплечья (файл LR2). Результаты исследования изображены на рис. 10.5.

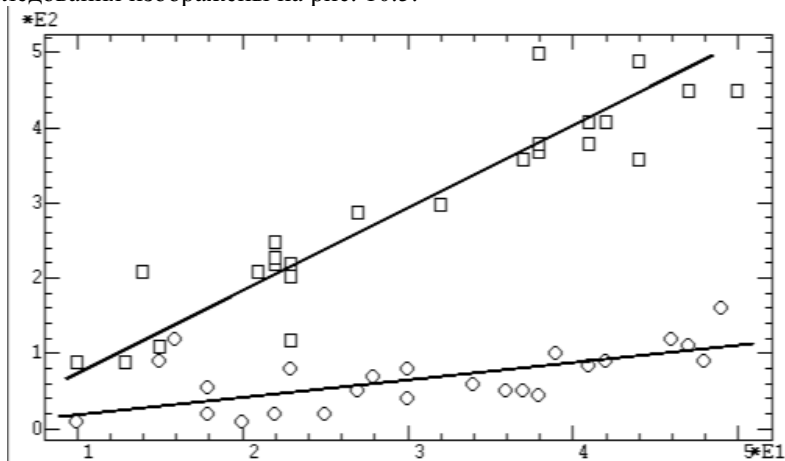


Рис. 10.5. Сила сжатия кисти (кг, ось  $Y$ ) в зависимости от поперечного сечения предплечья ( $\text{см}^2$ , ось  $X$ ) для здоровых людей (квадраты) и больных артритом (круги)

### Результаты:

СРАВНЕНИЕ ДВУХ РЕГРЕССИЙ. Файл: lr2.std

$T$ (параллельность)=7.83, Значимость=9.27E-7, степ.своб = 44

Гипотеза 1: <Есть различия между коэффициентами наклона>

$T$ (равенство средних)=19.8, Значимость=8.32E-10, степ.своб = 44

Гипотеза 1: <Есть различия в положении регрессионных прямых>

**В ы в о д ы:** Согласно результатам анализа выявлены различия как углов наклона, так и в сдвиге регрессионных зависимостей (оба уровня значимости близки к нулю).

## 10.3. Простая регрессия

### Пример 1

**З а д а ч а.** Рассмотрим данные Госкомстата СССР по средней урожайности зерновых культур (центнеры с гектара) с 1945 по 1989 годы (файл CORN). Попробуем проанализировать закономерности изменения этого кардинально важного для всей экономики и благосостояния страны показателя, чтобы иметь возможность строить достоверные прогнозы на будущее.

Для этого нам, прежде всего, полезно визуально изучить график данного временного ряда (рис. 10.20).

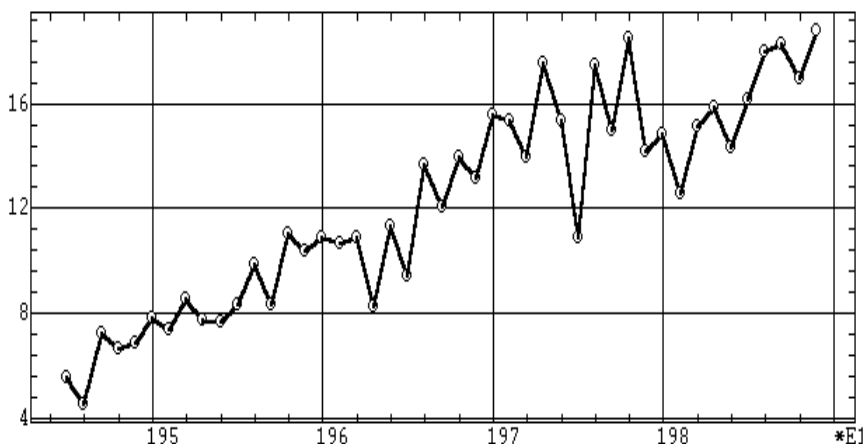


Рис. 10.20. Изменение урожайности зерновых в СССР с 1945 по 1989 гг.

Как легко заметить, в динамике урожайности преобладает линейно возрастающая тенденция с ежегодными нерегулярными колебаниями, поэтому естественным представляется описание этих данных линейной регрессионной моделью. Можно также отметить: два периода увеличения нестабильности в 1962–1966 гг. и в 1973–1981 гг., два периода стагнации в 1958–1962 гг. (эпоха позднего Хрущева) и в 1970–1982 гг. (эпоха позднего Брежнева), а также ускорение возрастания урожайности в 1963–1973 гг. (эпоха раннего Брежнева) и в 1983–1989 гг. (послебрежневская эпоха и эпоха раннего Горбачева). Вот сколько много полезных и ассоциативных исторических выводов можно сделать из простого просмотра исходных данных.

Перейдем собственно к регрессионному анализу, применив линейную модель и вычислив по ней прогностическое значение на 2000 год.

### Результаты:

ПРОСТАЯ РЕГРЕССИЯ. Файл: corn.std Переменные: data, зерно

Модель: линейная  $Y = a_0 + a_1 \cdot x$

Коэфф.	a0	a1
Значение	-529	0.275
Ст.ошиб.	36.3	0.0185
Значим.	0	0

Источник Сум.кв. Св Средн.кв. Св.

Регресс.	574	1	574
Остаточн	111	43	2.59
Вся	686	44	

Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим
0.91521	0.8376	0.83383	1.6094	222	0

Гипотеза1: <Регрессионная модель адекватна экспериментальн. данным>

data=2000, зерно=21.274

Хэксп	Yэксп	Yрегр	остаток	Ст.остат	Ст.ошиб	Довер.инт
1.95E3	5.6	6.14	-0.543	-0.341	1.68	3.34
1.95E3	4.6	6.42	-1.82	-1.14	1.67	3.33
1.95E3	7.3	6.69	0.607	0.381	1.67	3.32
1.95E3	6.7	6.97	-0.268	-0.169	1.66	3.31
1.95E3	6.9	7.24	-0.344	-0.216	1.66	3.31
1.95E3	7.9	7.52	0.381	0.24	1.66	3.3

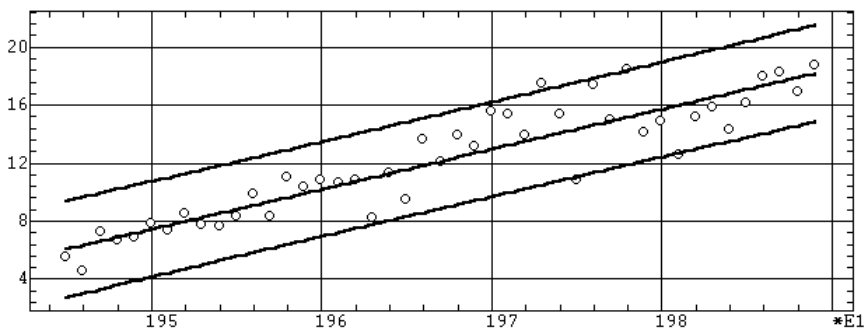
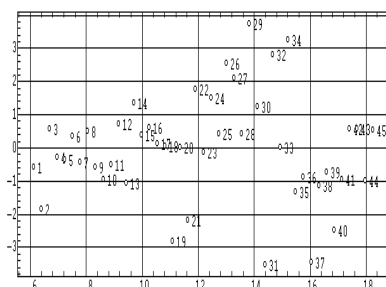
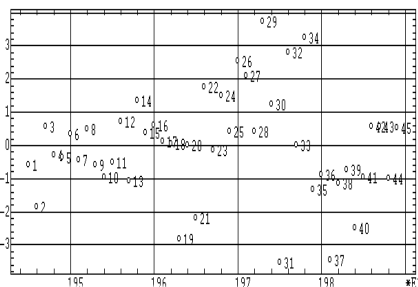


Рис. 10.21. Линейная регрессионная модель с зоной доверительного интервала



$a$   $b$   
 Рис. 10.22. Регрессионные остатки (по оси  $Y$ ):  
 $a$  — по годам;  $b$  — относительно регрессионных значений

**Обсуждение результатов.** Как следует из числовых результатов, линейная модель адекватна экспериментальным данным (значимость нулевой гипотезы близка к нулю). На регрессионном графике (рис. 10.21) лишь незначительное число экспериментальных точек выходит за доверительный интервал, а распределение остатков (рис. 10.22) достаточно однородно. Если сохранить остатки в электронной таблице, то при дальнейшем анализе корреляций и распределений легко выяснить, что их распределение не коррелировано как по  $X$ , так и по  $Y$  и нормально распределено, и это дополнительно подтверждает адекватность модели.

Мы получили по построенной модели для 2000 г. прогноз урожайности более 21 центнера с га, однако можно попытаться построить и детальный погодовой прогноз.

**Выдача прогноза:**

Хпрогн	Упрогн	Ст.ошиб	Довер.инт
1.99Е3	18.5	1.68	3.35
1.99Е3	18.8	1.69	3.36
1.99Е3	19.1	1.69	3.37
1.99Е3	19.3	1.7	3.38
1.99Е3	19.6	1.7	3.39
2Е3	19.9	1.71	3.4
2Е3	20.2	1.71	3.41
2Е3	20.4	1.72	3.42
2Е3	20.7	1.73	3.43
2Е3	21	1.73	3.45

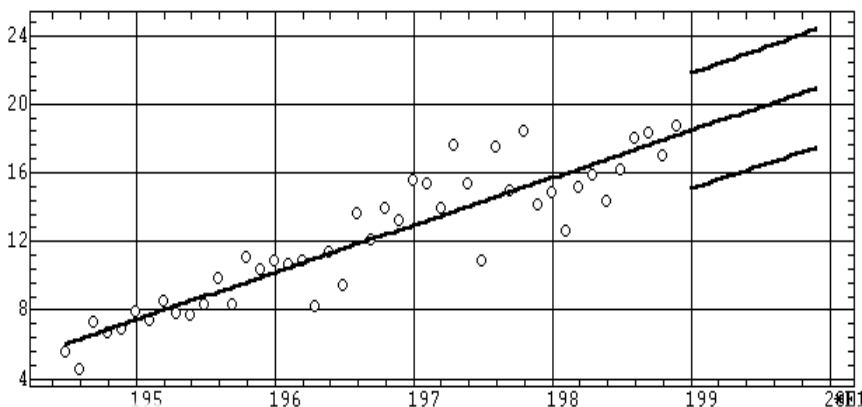


Рис. 10.23. Прогнозирование урожайности пшеницы с 1990 по 2000 гг. с зоной доверительного интервала прогноза индивидуальных значений

**Обсуждение:** Если сравнить среднюю урожайность зерновых в 1990—1996 гг. с нашим прогнозом, то выявленная линейная тенденция

продолжает сохраняться, а ежегодные колебания вполне укладываются в доверительный интервал прогноза.

Полученные результаты могут навести на крайне интересные геополитические выводы:

- 1) средняя тенденция увеличения урожайности дает некоторую надежду на оттягивание глобального голода и вымирания человечества;
- 2) отчаянные попытки «догнать и перегнать Америку» и внедрение кукурузы в 60–х годах, а также криминальный развал народного хозяйства в 90–х годах не смогли, к счастью, резко переломить общую урожайную тенденцию;
- 3) экспоненциальное увеличение научных знаний и технологий также не позволяет изменить стабильность коэффициента наклона, видимо, компенсируясь соразмерно прогрессирующей деградацией земель.

*Продолжение анализа.* Но не будем торопиться, а еще раз повнимательнее присмотримся к динамике роста урожайности (см. рис. 10.20). И зададимся вопросом: так ли она уж строго линейна, не видится ли в ней признаков замедления роста, особенно на последнем отрезке?

Для ответа на этот вопрос можно сделать следующее: разбить интервал наблюдения на четыре последовательных отрезка, подсчитать на каждом среднее значение и вычислить разность средних.

Для последовательных 11–летних периодов средствами описательной статистики мы получим следующие средние значения: 7,173, 10,47, 14,61, 15,93, а их разности 3,3, 2,14, 1,32 показывают очевидное замедление роста. Поэтому продолжим анализ с подбором более адекватной модели. После перебора возможных вариантов наиболее приемлемой следует признать параболическую модель.

### **Результаты:**

ПРОСТАЯ РЕГРЕССИЯ. Файл: corn.std Переменные: data, zerno

Модель: парабола  $Y = a_0 + a_1 \cdot x + a_2 \cdot x^2$

Козфф.	a0	a1	a2
Значение	-10445	10.36	-0.0025
Ст.ошиб.	6035.5	6.137	0.0015
Значим.	0.087	0.095	0.104

Источник	Сум.кв.др.	Степ.св	Средн.кв.др.
Регресс.	581.19	2	290.59
Остаточн	104.65	42	2.4917
Вся	685.84	44	

Множество R	R^2	R^2прив	Ст.ошиб.	F	Значим
0.92055	0.84741	0.84014	1.5785	116.6	0

Гипотеза 1: <Регрессионная модель адекватна эксперимент.данным>

data=2000, zerno=18.915

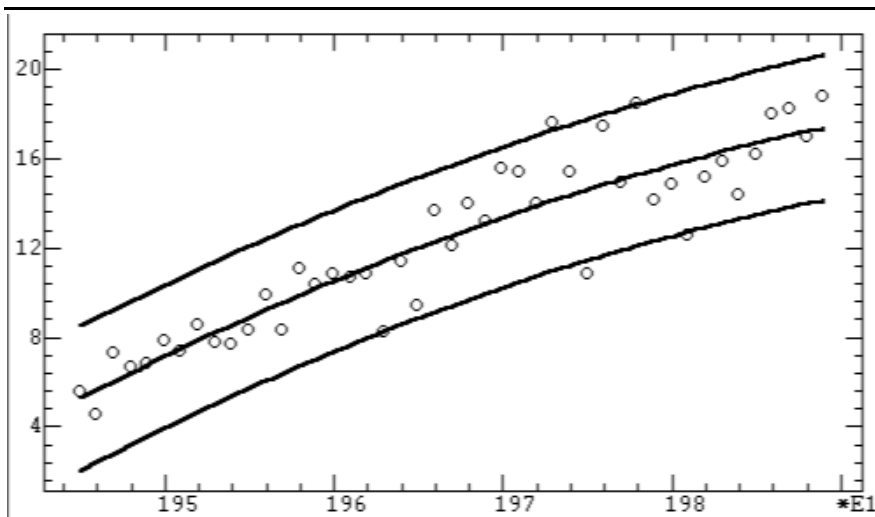


Рис. 10.24. Параболическая регрессионная модель с зоной доверительного интервала

**Обсуждение результатов.** Как легко заметить по итоговой статистике (более высокая множественная корреляция, меньше стандартная ошибка), параболическая модель является более адекватной для данного временного ряда. И эта модель в прогнозе дает заметно меньшую среднюю урожайность зерновых к 2000 году (19 вместо 21 ц/га).

## Пример 2

**З а д а ч а.** На химическом факультете МГУ учебная часть ведет детальный анализ успеваемости студентов в течение каждого семестра по следующим четырем разделам<sup>1</sup>: практикумы, коллоквиумы, синтезы, контрольные работы (рис. 10.25). В каждом разделе в течение семестра проводятся по 4–5 занятий и по каждому занятию для каждого студента выставляется балльная оценка. Затем оценки по каждому разделу суммируются, вычисляется общая сумма баллов, проводится экзамен, по которому выставляется 110–балльная оценка.

<sup>1</sup> Данные предоставлены профессором МГУ В.П. Зломановым

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1		Общая сумма	Практикум						Коллоквиумы						Синтезы						Контрольные						Экзамен
2			1	2	3	4	5	Сум	1	2	3	4	5	Сум	1	2	3	4	5	Сум	1	2	3	4	Сум		
3	Авраменко	228.3	6	4	7	7	6	30	15	15	17	17	17	81	4	4	4	4	4	20	12.0	8.5	19.5	9.5	49.5	48.3	
4	Батырев	130.1	6	3	7	5	4	25	14	14	13	13	12	66	4	4	4	4	4	20	6.5	9.0	2.5	0.5	18.5	0.6	
5	Година	236.1	6	4	7	7	6	30	20	20	20	20	20	100	4	4	4	4	4	20	18.5	24.5	19.5	22.0	84.5	1.6	
6	Ефимов Н.	18.8	6	4				10						0						0		2.5	2.0	3.5	8.0	0.8	
7	Иютси	195.0	6	4	7	7	6	30	15	16	18	20	20	89	4	4	4	4	4	20	8.0	8.0	20.5	18.0	54.5	1.5	
8	Коваленко	300.3	6	4	7	7	6	30	19	19	19	20	20	97	4	4	4	4	4	20	20.0	22.0	22.5	15.0	79.5	73.8	
9	Кружалова	300.0	6	4	7	7	6	30	18	19	19	20	20	96	4	4	4	4	4	20	23.0	19.0	25.0	12.5	79.5	74.5	
10	Ломако	232.8	6	4	7	7	6	30	19	20	19	20	20	98	4	4	4	4	4	20	24.5	17.0	22.0	18.0	81.5	3.3	
11	Меньшени	63.0	6	4				10	18					18	4					4	11.0	12.5	7.0		30.5	0.5	
12	Морозов Д.	241.0	6	4	7	7	6	30	20	20	19	20	19	98	4	4	4	4	4	20	25.0	23.5	22.5	20.5	91.5	1.5	
13	Новикова	293.2	6	4	7	7	6	30	19	19	19	19	20	96	4	4	4	4	4	20	21.0	14.5	23.0	10.0	68.5	78.7	
14	Носкова	295.2	6	4	7	7	6	30	18	19	19	19	19	94	4	4	4	4	4	20	16.0	16.0	21.5	19.0	72.5	79.2	
15	Попов М.В	291.0	6	4	7	7	6	30	19	18	18	18	19	92	4	4	4	4	4	20	23.0	11.5	20.5	15.5	70.5	78.5	
16	Простакова	237.5	6	4	7	7	6	30	19	20	20	20	18	97	4	4	4	4	4	20	22.0	24.0	25.0	15.5	86.5	4.0	
17	Пурецкий	310.8	6	4	7	7	6	30	19	19	19	20	19	96	4	4	4	4	4	20	24.0	22.5	18.0	14.5	79.0	85.8	
18	Рагузин	314.0	6	4	7	7	6	30	19	18	19	20	19	95	4	4	4	4	4	20	20.5	22.0	23.0	12.0	77.5	91.5	
19	Ракша	222.5	6	4	7	7	6	30	17	16	15	16	15	79	4	4	4	4	4	20	17.0	15.0	13.0	9.0	54.0	39.5	
20	Рассохина	286.5	6	4	7	7	6	30	18	19	19	20	19	95	4	4	4	4	4	20	18.5	18.0	22.5	17.0	76.0	65.5	
21	Самохин	302.7	6	4	7	7	6	30	18	18	18	18	17	89	4	4	4	4	4	20	19.5	17.0	22.5	14.0	73.0	90.7	
22	Смогоржев	185.5	6	4	7	7	6	30	16	17	18	17	18	86	4	4	4	4	4	20	8.0	12.5	14.5	13.0	48.0	1.5	
23	Стрелков	224.3	6	4	7	7	6	30	18	16	17	16	18	85	4	4	4	4	4	20	7.0	5.5	15.0	9.0	36.5	52.8	
24	Фролов	239.3	6	4	7	7	6	30	17	16	18	19	18	88	4	4	4	4	4	20	14.0	16.0	16.0	5.5	51.5	50.3	
25	Чистяков	179.0	6	4	7	7	6	30	16	14	16	17	17	80	4	4	4	4	4	20	5.0	8.0	11.0	5.5	29.5	19.5	
26	Шишин	235.6	6	4	7	7	6	30	19	20	19	20	20	98	4	4	4	4	4	20	23.5	22.5	20.0	20.3	86.3	1.3	

Рис. 10.25. Таблица учета семестровой успеваемости студентов одной группы химфака МГУ

Определяющими для успеваемости студента по семестру являются: сумма баллов за контрольные работы (максимум 100 баллов) и общая сумма баллов (максимум 255 баллов).

Преподавательский опыт говорит, что существует достаточно строгая связь между эффективностью работы студентов в семестре и их оценками на экзамене, который с учетом этой связи является достаточной формальностью. Поэтому учебная часть решила разработать метод априорного выставления экзаменационных оценок студентам по результатам их семестровой работы. Если же студент не согласен с априорной оценкой, то он может попытаться ее улучшить, пойдя на экзамен. Это позволило бы: 1) разгрузить преподавателей от излишней экзаменационной работы; 2) избавить студентов от дополнительного стресса, связанного с подготовкой и сдачей экзаменов.

**Визуальный анализ.** Сначала надо выяснить, действительно ли существует предполагаемая зависимость. Лучше всего это будет видно на диаграмме рассеяния. Возьмем для примера зависимость общего балла и экзаменационного балла по 246 студентам (переменные *Экз* и *Сумма* из файла CHMSTR).



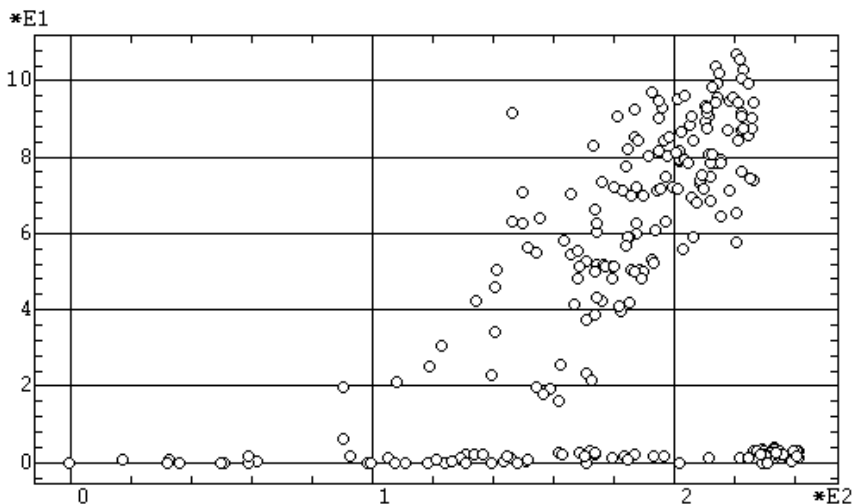


Рис. 10.26. Диаграмма рассеяния экзаменационный балл =  $f(\text{сумма баллов за семестр})$

**В ы в о д ы.** С некоторым упрощением подходу к данной задаче будем рассуждать следующим образом. Как следует из диаграммы (рис. 10.26), в области ненулевых экзаменационных оценок (по оси ординат) наблюдается близкая к линейной зависимость между анализируемыми показателями. Однако в исходных данных явно присутствует инородная и довольно большая группа студентов, имеющих нулевые экзаменационные оценки.

Выяснить истинную причину такого положения можно только при более глубоком изучении реальной ситуации. То есть здесь необходим переход от формального анализа к предметному анализу. В результате выясняется, что в эту инородную группу входят четыре категории студентов: а) не сдавшие экзамен по причине плохих знаний (пропустившие много семестровых занятий (от 0 до 100 баллов по оси абсцисс); б) не сдавшие экзамен по причине экзаменационного волнения (от 100 до 200 баллов по оси абсцисс); в) не пришедшие на экзамен; г) получившие «автомат» за отличную работу в семестре (от 200 до 250 баллов по оси абсцисс).

**Преобразование данных.** Присутствие таких инородных данных может сильно исказить результаты анализа, поэтому необходимо их удалить. Проще всего это сделать, выполнив операцию *сортировки* из Блока преобразований (разд. 3.4) для переменных *Экз* и *Сумма* по возрастанию значений переменной *Экз*, после чего удалить из матрицы данных все строки, содержащие малые значения переменной *Экз*. В

результате останутся данные по 156 студентам (переменные *Экз1* и *Сумма1* из файла CHMSTR).

**А н а л и з.** Рассчитаем для рафинированных таким образом данных регрессионную модель, по которой вычислим экзаменационные оценки для значений 150 и 200 суммы экзаменационных баллов:

ПРОСТАЯ РЕГРЕССИЯ. Файл: chmstr.std Переменные: сумма1, экз1

Модель: линейная  $Y = a_0 + a_1 \cdot x$

Коэфф.	$a_0$	$a_1$
Значение	-49.83	0.6207
Ст.ошиб.	8.434	0.04386
Значим.	4.511E-6	3.98E-10

Источник	Сум.кв.др.	Степ.св	Средн.кв.др.
Регресс.	4.438E4	1	4.438E4
Остаточн	3.412E4	154	221.6
Вся	7.85E4	155	

Множеств R	$R^2$	$R^2_{\text{прив}}$	Ст.ошиб.	F	Значим
0.75189	0.56534	0.56252	14.885	200.3	2.45E-16

Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>

сумма1=150,  $Y=43.27$

сумма1=200,  $Y=74.31$

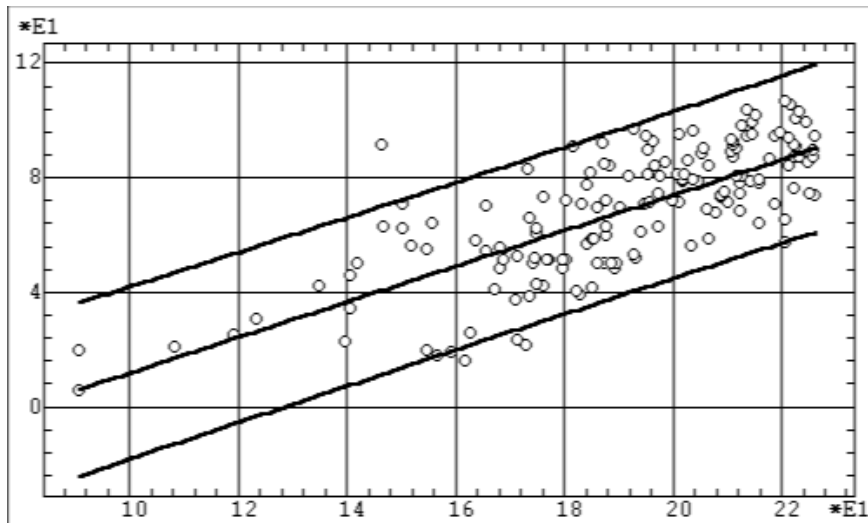


Рис. 10.27. Линейная регрессия экзаменационный балл =  $f(\text{сумма баллов за семестр})$  с доверительным интервалом

**В ы в о д ы:**

1. Линейная регрессионная модель адекватна экспериментальным данным на пренебрежительно малом уровне значимости нулевой гипотезы.

2. Модель позволяет рассчитать априорные экзаменационные оценки для студентов по сумме их семестровых данных. Так при сумме баллов 200 студенту может быть предложена экзаменационная оценка в 74

балла, что в пересчете на 5-бальную систему от максимума в 110 баллов составляет  $74/110 \cdot 5 = 3,364$ .

**Анализ остатков.** Однако такое предложение (вывод 2) будет статистически неполноценным, поскольку кроме регрессионной прямой надо учитывать еще и статистический разброс оценок. Поэтому надо продолжить регрессионный анализ: 1) выполнить анализ регрессионных остатков с построением их графика (рис. 10.28); 2) сохранить остатки с графика в электронной таблице нажатием на инструментальную кнопку «СохрГраф»; 3) вычислить описательную статистику для остатков; 4) при вычислении априорной экзаменационной оценки корректировать ее с учетом вычисленного стандартного отклонения регрессионных остатков.

### Результаты:

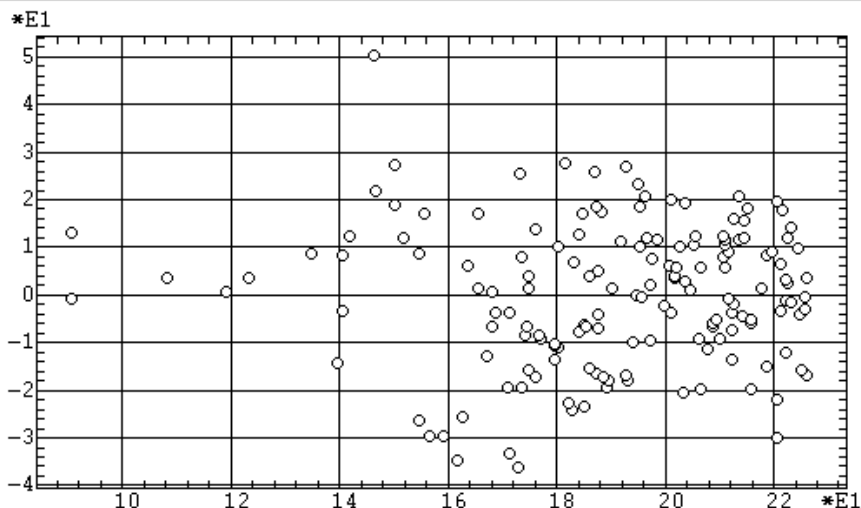


Рис. 10.29. Диаграмма рассеяния регрессионных остатков по значениям независимой переменной Сум

ОПИСАТЕЛЬНАЯ СТАТИСТИКА. Файл: chmstr.std

Перемен.	Размер	← Диапазон →	Среднее	Ошибка	Дисперс	Ст.откл	Сумма
x8	156	-36.7 50.2	-1.3E-8	1.18	220.1	14.84	-2E-6

**В ы в о д ы:** Стандартное отклонение регрессионных остатков в пересчете на 5-бальную систему составляет  $14,84/110 \cdot 5 = 0,6745$ . Поэтому при суммарном балле, равном 200, экзаменационная оценка может варьироваться от  $3,364 - 0,6745$  до  $3,364 + 0,6745$  (с учетом того, что диапазон одного стандартного отклонения включает 68% популяции). При выборе из этого диапазона следует учитывать действие случайных факторов, ведущих к снижению оценки на экзамене: влияние стресса, неудачного вопроса в билете, усталость преподавателя и т. п. Поэтому

представляется разумным предлагать студентам оценки из верхней части указанного диапазона.

**Обратная задача.** Кроме вышерассмотренной для преподавателя актуальна и обратная задача: иметь таблицу, в которой для проходных экзаменационных баллов были бы даны соответствующие им диапазоны суммарных баллов. Для этого нужно по вышерассмотренной схеме произвести обратный регрессионный анализ для зависимости суммы баллов от экзаменационной оценки. При этом вычислим суммы баллов для оценок 5 (110 баллов), 4 (88 баллов), 3 (66 баллов), 2 (44 балла) и оценки 74 балла для сравнения с ранее полученными результатами.

### Результаты:

ПРОСТАЯ РЕГРЕССИЯ. Файл: chmstr.std Переменные: экз1, суммал

Модель: линейная  $Y = a_0 + a_1 \cdot x$

Коэфф.	a0	a1
Значение	128.1	0.9108
Ст.ошиб.	4.629	0.06436
Значим.	1.01E-12	3.98E-10

Источник Сум.кв. Степ.св Средн.кв.др.

Регресс. 6.513E4 1 6.513E4

Остаточн 5.007E4 154 325.1

Вся 1.152E5 155

Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим
------------	-----	---------	----------	---	--------

0.75189	0.56534	0.56252	18.032	200.3	2.45E-16
---------	---------	---------	--------	-------	----------

Гипотеза 1: <Регрессионная модель адекватна эксперимент.данным>

экз1=110, суммал=228.3

экз1=88, суммал=208.3

экз1=66, суммал=188.2

экз1=44, суммал=168.1

экз1=74, суммал=195.5

### Выводы:

1. Линейная регрессионная модель адекватна экспериментальным данным.
2. Определены средние значения суммы баллов для основных экзаменационных оценок: 228, 208, 188, 168, 195.
3. Полученное значение 195,5 суммы баллов для экзаменационной оценки 74 немного не совпадает со значением 200, использованным выше в прямой регрессии (расхождение составляет менее 2,5%). Это является следствием несовпадения прямой и обратной регрессии, обсуждаемого во введении к данной главе.
4. Если продолжить анализ регрессионных остатков по вышерассмотренной схеме, то можно определить стандартное отклонение регрессии, равное 17,97.

**Направления дальнейшего анализа.** Имеющиеся данные предоставляют богатейший материал для разнообразных дальнейших

исследований. Сформулируем только одну из интересных в этом плане задач.

Различные группы студентов ведут разные преподаватели. У каждого из них имеются свои правила преподавания и выставления оценок, что сказывается как на знаниях студентов, так и на полученных ими баллах. Сделав допущение, что группы студентов примерно равны по средним способностям, можно для каждого преподавателя вычислить средний бал выставляемых ими оценок, а также выявить их попарные статистические различия и группировки. Зная это, можно оценить качество преподавания предмета различными преподавателями, сравнивая баллы, выставленные их студентам другими преподавателями. Все это может быть плодотворным материалом для организационных и методических выводов по коррекции учебного процесса с целью повышения уровня преподавания, а также ликвидации систематического завышения или занижения оценок.

## **10.4. Множественная линейная регрессия**

---

### **Пример**

**Задача.** Мировая статистика накопила богатейший материал по множеству экономических и социальных показателей развития различных стран. Эти данные, в частности, могут быть полезны для исследования разного рода зависимостей и прогнозирования. Многие из таких зависимостей носят линейный характер, или же могут быть приближены к таковому посредством простых алгебраических преобразований исходных показателей.

Однако линейное моделирование осмыслено не для всего множества показателей, а только для тех из них, которые взаимосвязаны. Поэтому на первом этапе такого исследования необходимо вычислить кросс-корреляционную матрицу и отобрать для дальнейшего анализа группы показателей, взаимно связанных большими корреляциями.

На таком предварительном этапе для данного примера из данных по 108 государствам мы отобрали следующие взаимосвязанные показатели<sup>1</sup> (файл MLR): средняя продолжительность жизни, детская смертность, потребление калорий, уровень рождаемости, уровень смертности, количество детей, взаимно связанными высокими корреляциями:

---

<sup>1</sup> Данные из архива SPSS.

ПАРАМЕТРИЧЕСКАЯ КОРРЕЛЯЦИЯ. Файл: mlr.std

Корреляционная матрица

	ДЛ.ЖИЗНИ	ДЕТСМЕРТ	КАЛОРИИ	УР.РОЖД	УР.СМЕРТ
ДЕТСМЕРТ	-0.946				
КАЛОРИИ	0.765	-0.777			
УР.РОЖД	-0.829	0.867	-0.762		
УР.СМЕРТ	-0.756	0.646	-0.355	0.432	
КОЛДЕТЕЙ	-0.809	0.845	-0.696	0.974	0.48
Критическое значение=0.224					
Число значимых коэффициентов=15 (100%)					

Выявление изменений в средней продолжительности жизни требует достаточно длительного времени наблюдения. Изменения же в других показателях могут быть отслежены значительно быстрее. Поэтому актуальной является задача предсказания изменения продолжительности жизни при изменении других, связанных с ней показателей. На первом этапе такого исследования предположим, что продолжительность жизни линейно зависит от других социальных показателей, и для решения поставленной задачи рассчитаем линейную многопараметрическую регрессионную модель. После этого осуществим предсказание длительности жизни для избранного набора значений других показателей.

### Результаты:

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ. Файл: mlr.std

Y=ДЛ.ЖИЗНИ	ДЕТСМЕРТ	КАЛОРИИ	УР.РОЖД	УР.СМЕРТ	КОЛДЕТЕЙ	
Коэфф.	a0	a1	a2	a3	a4	a5
Значение	74.9	-0.123	0.00247	-0.332	-0.787	1.3
Ст.ошиб.	3.19	0.0194	0.000846	0.121	0.0914	0.68
Значим.	4.68E-11	3.79E-6	0.00496	0.00787	1.77E-7	0,0564

Источник Сум.кв. Св. Степ.св Средн.кв. Св.

Регресс. 7.19E3 5 1.44E3

Остаточн 375 69 5.43

Вся 7.56E3 74

Множеств R	R <sup>2</sup>	R <sup>2</sup> прив	Ст.ошиб.	F	Значим
0.97491	0.95046	0.94687	2.3305	265	4.74E-12

Гипотеза 1: &lt;Регрессионная модель адекватна эксперименталь.данным&gt;

ДЕТСМЕРТ=20, КАЛОРИИ=2Е3, УР.РОЖД=10, УР.СМЕРТ=5, КОЛДЕТЕЙ=3

ДЛ.ЖИЗНИ=74

**Обсуждение:** Как можно заметить, построенная линейная модель адекватна экспериментальным данным (уровень значимости гипотезы о равенстве нулю коэффициента множественной корреляции близок к нулю). Однако, рассматривая значимости регрессионных коэффициентов при различных показателях, можно заметить, что количеством детей в дальнейшем можно пренебречь, поскольку значимость гипотезы о равенстве его коэффициента  $a_5$  нулю ( $P=0,0564$ ) выше критического уровня. Поэтому повторим анализ с исключением этого показателя.

### Результаты:

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ. Файл: mlr.std

Y=ДЛ.ЖИЗНИ	ДЕТСМЕРТ	КАЛОРИИ	УР.РОЖД	УР.СМЕРТ	
Коэфф.	a0	a1	a2	a3	a4

Значение	72.3	-0.126	0.00292	-0.118	-0.734
Ст.ошиб.	2.94	0.0197	0.000828	0.0481	0.0889
Значим.	3.33E-11	3.33E-6	0.00109	0.0161	2.59E-7

Источник Сум.квадр. Степ.св Средн.квадр.

Регресс.	7.17E3	4	1.79E3
Остаточн	395	70	5.64
Вся	7.56E3	74	

Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим
0.97356	0.94782	0.94484	2.3745	318	4.43E-12

Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>

ДЕТСМЕРТ=20, КАЛОРИИ=2Е3, УР.РОЖД=10, УР.СМЕРТ=5, ДЛ.ЖИЗНИ=70.7

**Обсуждение:** Уточненная модель осталась практически на том же уровне адекватности (при некотором увеличении значения  $F$ -статистики), но можно заметить, что при тех же значениях параметров она предсказывает несколько меньшую продолжительность жизни (70,7 лет вместо 74). Теперь можно произвести анализ регрессионных остатков и построение результирующих графиков.

Хэксп	Уэксп	Урегр	остаток	Ст.остат	Ст.ошиб	Довер.инт
7.3	74	73.1	0.902	0.391	2.36	4.64
6.7	73	72.1	0.862	0.373	2.36	4.64
7.2	74	70.9	3.05	1.32	2.36	4.64
25.6	68	69.2	-1.16	-0.503	2.35	4.62
106	53	52.6	0.424	0.184	2.38	4.68
75	59	57.8	1.22	0.53	2.35	4.62
39.3	60	64.6	-4.6	-1.99	2.34	4.61
66	57	62.9	-5.88	-2.55	2.34	4.62
118	47	45.3	1.71	0.742	2.39	4.71
105	46	44	1.96	0.848	2.38	4.68

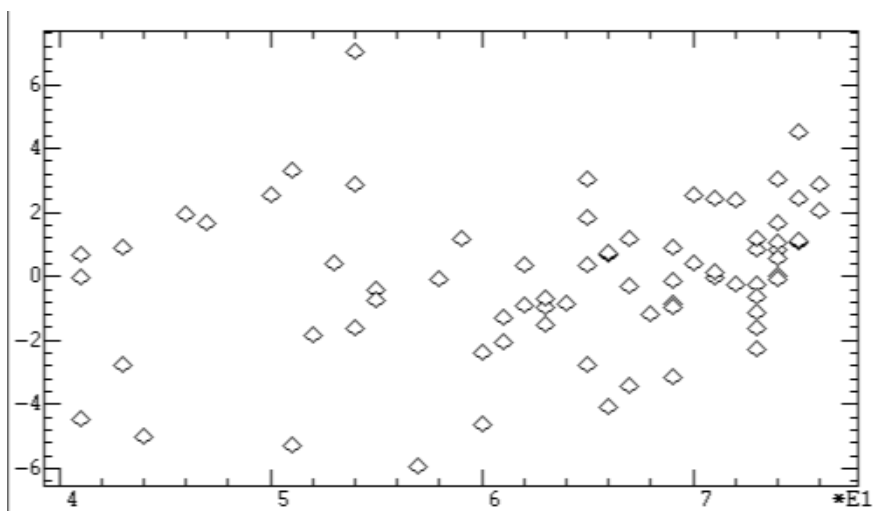


Рис. 10.32. Регрессионные остатки (по оси Y) относительно регрессионных значений (по оси X)

*Обсуждение:* Визуальное изучение распределения регрессионных остатков (рис. 10.32) не выявляет какой-либо зависимости между ними. Заметна только более плотная концентрация в области больших длительностей жизни. Это говорит о том, что страны с меньшей длительностью жизни более сильно различаются между собой.

Изучение регрессионных графиков показывает следующее: 1) продолжительность жизни практически линейно связана с детской смертностью (рис. 10.33, *а*), что вполне соответствует смысловой взаимосвязи этих показателей; 2) хорошая линейная связь прослеживается и с уровнем смертности (рис. 10.33, *з*); 3) связи с уровнем рождаемости и особенно с потреблением калорий носят нелинейный характер, поэтому для большей адекватности в отношении прогноза возможно будет оправдано введение в модель нелинейных членов по этим показателям (см. разд. 10.6).

Продолжение анализа данного примера см. в следующем разделе.



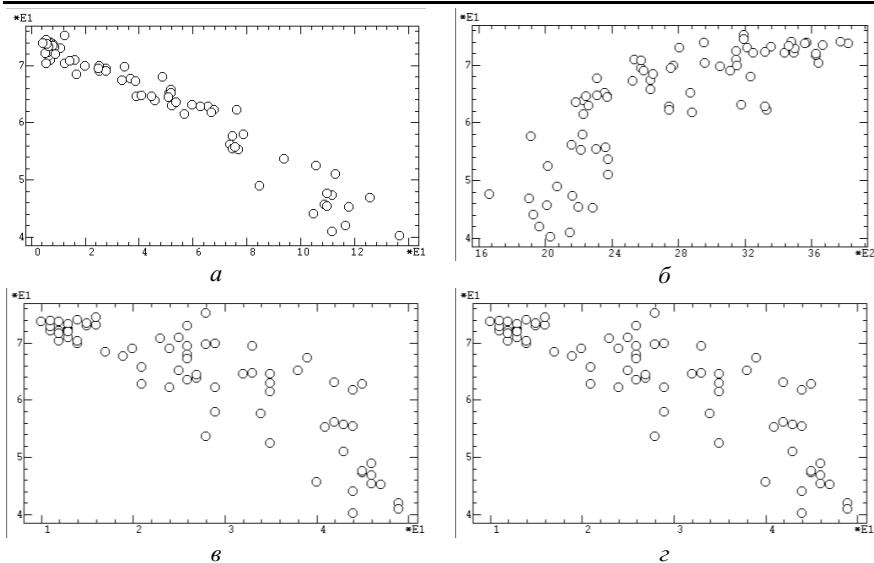


Рис. 10.33. Распределения регрессионных значений (по оси  $Y$ ) относительно исходных показателей (по оси  $X$ ):

$a$  — детской смертности;  $б$  — калорий;  
 $в$  — уровня рождаемости;  $г$  — уровня смертности

## 10.5. Пошаговая регрессия

### Примеры

**З а д а ч а.** Мы продолжим анализ данных мировой социальной статистики из примера к разд. 10.4 с целью определить методом шаговой регрессии необходимый и достаточный набор переменных, объясняющих показатель средней продолжительности жизни.

#### Результаты:

ПОШАГОВАЯ РЕГРЕССИЯ. Файл: mlr.std

\*\*\* Метод включения. Шаг No.1, введена переменная:ДЕТСМЕРТ

Коэфф.	a0	a1
Значение	75.5	-0.247
Ст.ошиб.	0.603	0.00992
Значим.	9.65E-14	2.6E-11

Источник	Сум.квадр.	Степ.св	Средн.квадр.
Регресс.	6.77E3	1	6.77E3
Остаточн	797	73	10.9
Вся	7.56E3	74	

Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим
0.94583	0.8946	0.89316	3.3047	620	3.24E-14

Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>  
Измен.R^2 F Значим

0.895	620	2.6E-11					
----- Переменные в уравнении -----							
Переменн.	Коэфф.В	Ст.ош.В	Бета	F	Значим		
ДЕТСМЕРТ	-0.247	0.00992	-0.946	620	2.6E-11		
----- Переменные не в уравнении -----							
Переменн.	Коэфф.В	Ст.ош.В	Бета	F	Значим	Частн. R	Толер.
КАЛОРИИ	0.00136	0.00107	0.0763	1.61	0.206	0.148	0.396
УР. РОЖД	-0.0308	0.0621	-0.038	0.246	0.627	0,0584	0.249
УР. СМЕРТ	-0.568	0.0924	-0.249	37.8	4.9E-6	0.587	0.583
КОЛДЕТЕЙ	-0.178	0.371	-0.034	0.23	0.639	0,0564	0.287
*** Метод включения. Шаг No.2, введена переменная: УР. СМЕРТ							
Коэфф.	a0	a1	a2				
Значение	79.1	-0.205	-0.568				
Ст.ошиб.	0.768	0.0106	0.0924				
Значим.	1.45E-13	1.5E-10	4.95E-6				
Источник	Сум.кв.др.	Степ.св	Средн.кв.др.				
Регресс.	7.04E3	2	3.52E3				
Остаточн	523	72	7.26				
Вся	7.56E3	74					
Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим		
0.96481	0.93086	0.92894	2.6951	485	6.77E-12		
Гипотеза 1: <Регрессионная модель адекватна эксперименталь. данным>							
Измен. R^2	F	Значим					
0.0363	37.8	4.95E-6					
----- Переменные в уравнении -----							
Переменн.	Коэфф.В	Ст.ош.В	Бета	F	Значим		
ДЕТСМЕРТ	-0.205	0.0106	-0.785	374	1.5E-10		
УР. СМЕРТ	-0.568	0.0924	-0.249	37.8	4.95E-6		
----- Переменные не в уравнении -----							
Переменн.	Коэфф.В	Ст.ош.В	Бета	F	Значим	Частн. R	Толер.
КАЛОРИИ	0.00332	0.000839	0.187	15.7	0.00038	0.425	0.359
УР. РОЖД	-0.152	0,0508	-0.187	8.91	0.00415	0.334	0.221
КОЛДЕТЕЙ	-0.489	0.302	-0.094	2.63	0.106	0.189	0.279
*** Метод включения. Шаг No.3, введена переменная: КАЛОРИИ							
Коэфф.	a0	a1	a2	a3			
Значение	68.9	-0.159	-0.675	0.00332			
Ст.ошиб.	2.68	0.0151	0.0885	0.000839			
Значим.	2.39E-11	2.4E-8	5.6E-7	0.000384			
Источник	Сум.кв.др.	Степ.св	Средн.кв.др.				
Регресс.	7.14E3	3	2.38E3				
Остаточн	428	71	6.03				
Вся	7.56E3	74					
Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим		
0.97127	0.94336	0.94096	2.4565	394	4.55E-12		
Гипотеза 1: <Регрессионная модель адекватна эксперименталь. данным>							
Измен. R^2	F	Значим					
0.0125	15.7	0.000384					
----- Переменные в уравнении -----							
Переменн.	Коэфф.В	Ст.ош.В	Бета	F	Значим		
ДЕТСМЕРТ	-0.159	0.0151	-0.609	111	2.4E-8		
УР. СМЕРТ	-0.675	0.0885	-0.297	58.2	5.6E-7		
КАЛОРИИ	0.00332	0.000839	0.187	15.7	0.000384		
----- Переменные не в уравнении -----							
Переменн.	Коэфф.В	Ст.ош.В	Бета	F	Значим	Частн. R	Толер.
УР. РОЖД	-0.118	0.0481	-0.145	5.99	0.0161	0.281	0.212
КОЛДЕТЕЙ	-0.411	0.277	-0.0789	2.2	0.138	0.175	0.278
*** Метод включения. Шаг No.4, введена переменная: УР. РОЖД							
Коэфф.	a0	a1	a2	a3	a4		
Значение	72.3	-0.126	-0.734	0.00292	-0.118		
Ст.ошиб.	2.94	0.0197	0.0889	0.000828	0.0481		
Значим.	3.33E-11	3.33E-6	2.59E-7	0.00109	0.0161		
Источник	Сум.кв.др.	Степ.св	Средн.кв.др.				
Регресс.	7.17E3	4	1.79E3				
Остаточн	395	70	5.64				
Вся	7.56E3	74					
Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим		

0.97356	0.94782	0.94484	2.3745	318	4.43E-12
Гипотеза 1: <Регрессионная модель адекватна эксперименталь. данным>					
Измен. R^2	F	Значим			
0.00446	5.99	0.0161			

----- Переменные в уравнении -----							
Переменн.	Коэфф. В	Ст. ош. В	Бета	F	Значим		
ДЕТСМЕРТ	-0.126	0.0197	-0.484	41	3.33E-6		
УР. СМЕРТ	-0.734	0.0889	-0.323	68.3	2.59E-7		
КАЛОРИИ	0.00292	0.000828	0.164	12.4	0.00109		
УР. РОЖД	-0.118	0.0481	-0.145	5.99	0.0161		
----- Переменные не в уравнении -----							
Переменн.	Коэфф. В	Ст. ош. В	Бета	F	Значим	Частн. R	Толер.
КОЛДЕТЕЙ	1.3	0.68	0.25	3.67	0,0564	0.225	0.042

**Обсуждение:** На последнем шаге процедуры формируется модель, включающая четыре переменные за исключением количества детей. Это подтверждает вывод, сделанный в примере к разд. 10.4, что этот показатель незначим для предсказания.

Отметим, что при пошаговой регрессии сравнительную значимость объясняющих переменных представляется возможным количественно оценить толерантностью, определяющей последовательность включения в модель.

Продолжение анализа данного примера см. в следующем разделе.

## 10.6. Общая регрессия

### Пример

**З а д а ч а.** Мы продолжим анализ данных мировой социальной статистики из примера к разд. 10.4. Там мы обнаружили нелинейный характер зависимости средней продолжительности жизни от количества потребляемых калорий и уровня рождаемости (рис 10.26). Попробуем учесть эти нелинейности введением в модель двух квадратичных членов. Формула нелинейной модели, вводимая в бланк формул, будет выглядеть следующим образом:

$$a_0 + a_1 * \text{ДЕТСМЕРТ} + a_2 * \text{КАЛОРИИ} + a_3 * \text{КАЛОРИИ}^2 + a_4 * \text{УР. РОЖД} + a_5 * \text{УР. РОЖД}^2 + a_6 * \text{УР. СМЕРТ}$$

Как и в примере раздела 10.4 вычислим прогноз при тех же значениях параметров модели.

### Результаты:

ОБЩАЯ (+НЕЛИНЕЙНАЯ) РЕГРЕССИЯ. Файл: mlr.std

Модель:

ДЛ. ЖИЗНИ =  $a_0 + a_1 * \text{ДЕТСМЕРТ} + a_2 * \text{КАЛОРИИ} + a_3 * \text{КАЛОРИИ}^2 + a_4 * \text{УР. РОЖД} + a_5 * \text{УР. РОЖД}^2 + a_6 * \text{УР. СМЕРТ}$

Коэфф.	a0	a1	a2	a3	a4	a5	a6
Значение	88	-0.1	0.0035	-3.7E-7	-1.04	0.0151	-1.13

Ст.ошиб.	9.42	0.0194	0.00623	1.1E-6	0.236	0.00377	0.128
Значим.	8.2E-8	2.95E-5	0.583	0.737	0.00014	0.000352	1.46E-7

Источник	Сум.квадр.	Степ.св	Средн.квадр.
Регресс.	7.25E3	6	1.21E3
Остаточн	318	68	4.67
Вся	7.56E3	74	

Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим
0.97876	0.95797	0.95426	2.1621	258	3.99E-12

Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>

ДЕТСМЕРТ=20, КАЛОРИИ=2Е3, УР.РОЖД=10, УР.СМЕРТ=5, ДЛ.ЖИЗНИ =76.9

**В ы в о д ы:** Модель адекватна исходным данным, однако коэффициенты  $a_2, a_3$  при показателе потребления калорий не отличны от нуля, поэтому скорректируем модель, удалив наиболее недостоверный квадратичный член, и повторим анализ.

### Результаты:

ОБЩАЯ (+НЕЛИНЕЙНАЯ) РЕГРЕССИЯ. Файл: mlr.std

Модель:

ДЛ.ЖИЗНИ= $a_0 + a_1 \cdot \text{ДЕТСМЕРТ} + a_2 \cdot \text{КАЛОРИИ} + a_3 \cdot \text{УР.РОЖД} + a_4 \cdot \text{УР.РОЖД}^2 + a_5 \cdot \text{УР.СМЕРТ}$

Коэфф.	a0	a1	a2	a3	a4	a5
Значение	90.6	-0.101	0.00142	-1.02	0.0148	-1.13
Ст.ошиб.	5.25	0.0189	0.000836	0.228	0.00365	0.127
Значим.	3.99E-10	2.03E-5	0.0896	0.000116	0.000309	1.25E-7

Источник	Сум.квадр.	Степ.св	Средн.квадр.
Регресс.	7.25E3	5	1.45E3
Остаточн	318	69	4.61
Вся	7.56E3	74	

Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим
0.97873	0.9579	0.95485	2.1482	314	3.29E-12

Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>

ДЕТСМЕРТ=20, КАЛОРИИ=2Е3, УР.РОЖД=10, УР.СМЕРТ=5, ДЛ.ЖИЗНИ=77

**В ы в о д ы:** Как можно заметить, значимость гипотезы о равенстве нулю коэффициента при потреблении калорий при исключении квадратичного члена резко снизилась до уровня, близкого к критическому (0,0896 против 0,583). Можно сделать вывод, что нелинейность по этому показателю определяется зависимостью ее от уровня смертности: чем более калорийно питание, тем меньше смертность. Почему возможна такая связанная нелинейная зависимость по двум показателям?

Представим себе мысленно более простой и наглядный пример: двумерную регрессионную плоскость в трехмерном пространстве  $y-x_1-x_2$ , причем эта плоскость линейна по координате  $x_1$ , но изогнута по координате  $x_2$ . Далее представим, что на этой плоскости рассыпаны наши экспериментальные точки, однако эти точки рассыпаны не хаотично, а группируются вдоль биссектрисы угла  $x_1-x_2$ . Тогда проекции точек как на координатную плоскость  $y-x_1$ , так и на координатную плоскость  $y-x_2$  будут изогнуты, несмотря на то, что сама регрессионная плоскость изогнута только по одной координате.

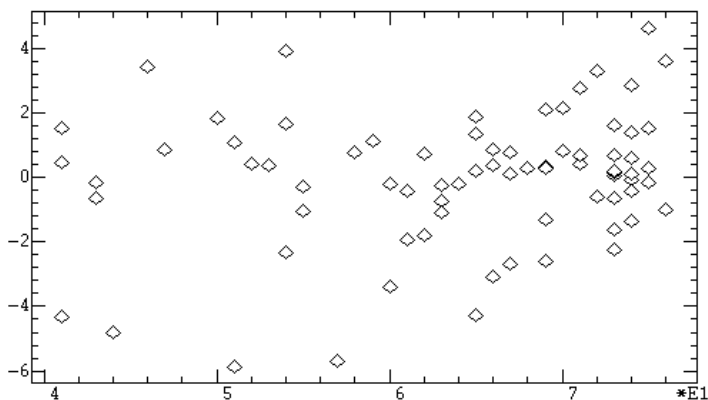


Рис. 10.38. Регрессионные остатки по оси  $Y$  относительно регрессионных значений по оси  $X$

Проведя далее анализ регрессионных остатков и построив график их распределения (рис. 10.38) можно заметить уменьшение диапазона разброса остатков (по оси ординат) по сравнению с линейной моделью (см. рис. 10.32). Это является дополнительным свидетельством большей адекватности нелинейной модели.

Учитывая близость к критической значимости коэффициента  $a_2$  можно оставить его в модели и использовать эту окончательную модель для прогнозирования средней продолжительности жизни. При введенных параметрах прогноз дает нам длительность жизни в 77 лет.

## **МНОГОМЕРНЫЕ МЕТОДЫ**

### **11.1. Факторный анализ**

#### **Пример 1**

**Задача.** Данный пример выполнен в психодиагностической методике «Сказка» для определения уровня развития детей<sup>1</sup>, в которой ребенку предлагается оценить по трехбалльной шкале «Да»–«Не знаю»–«Нет» (выражаются баллами 1, 0, –1) черты характера и личностные особенности (свойства) известных сказочных героев: Айболит, Буратино, Кот в сапогах, Снежная королева, Карабас–Барабас, Карлсон, Мальвина и Пьеро.

<sup>1</sup> Петренко В.Ф. Психосемантика сознания. М.: МГУ, 1988.

**Важное примечание.** Здесь мы имеем очень важный дидактический пример перехода от ранговой шкалы измерений (1, 0, -1) к метрической, более приемлемой для факторного анализа. Такой переход осуществляется посредством операции усреднения ответов многих респондентов.

В данной методике анализ факторной структуры в пространстве свойств героев позволяет выявить структуру сознания ребенка и судить об уровне сформированности и усвоенности общественных стереотипов и норм поведения. Для этого индивидуальные результаты ответов ребенка сравниваются с факторными свойствами нормативно-усредненной матрицы, полученной из ответов преподавателей начальной школы (табл. 11.1, файл TALES, в котором переменные соответствуют строкам табл. 11.1).

*Таблица 11.1. Характеристики сказочных персонажей*

Характер	Айболит	Кот	Карлсон	СнКор	Кар-Бар	Бурат	Мальвина	Пьер
ВернДруг	1.00	0.87	0.47	-0.98	-0.80	0.80	0.10	0.93
Смелый	0.93	0.87	0.20	0.30	-0.13	0.87	-0.33	0.13
Красивый	-0.27	0.53	-0.20	1.00	-0.80	-0.33	0.93	0.47
Добрый	1.00	0.73	0.60	-1.00	-1.00	0.80	0.21	0.87
Хитрый	-0.87	0.9	0.50	0.60	0.50	0.40	-0.93	-0.93
Жадный	-1.00	-0.67	0.53	0.98	1.00	-0.80	-0.87	-1.00
Плакса	-1.00	-0.95	-0.93	-0.67	0.00	-0.95	0.30	0.90
Умный	1.00	0.87	0.58	0.10	-0.27	0.43	0.44	0.37
Ябеда	-1.00	-1.00	-0.67	-1.00	0.78	-0.91	0.87	-0.83
Воспитан	0.93	0.93	-0.80	0.00	-0.90	-0.70	1.00	0.87
Хвостун	-1.00	0.27	1.00	0.33	0.60	0.77	0.00	-1.00
Умелый	1.00	0.93	0.20	0.03	-0.47	0.10	0.40	0.23
Драчун	-0.98	0.53	0.20	0.70	1.00	0.73	-1.00	-0.98
Шалун	-0.97	0.27	1.00	-0.87	0.00	1.00	-1.00	-0.80
Веселый	0.63	0.87	1.00	-1.00	-0.93	1.00	0.01	-1.00

Ниже производится анализ этой матрицы усредненных ответов преподавателей.

### **Результаты:**

```

ФАКТОРНЫЙ АНАЛИЗ.  файл: tale1.txt
Переменная    Среднее    Ст.отклон.
Верн.друг    0.299      0.791
Смелый        0.322      0.51
Красивый      0.166      0.655
Добрый        0.276      0.821
Хитрый        0.0213     0.785
Жадный        -0.229     0.9
Плакса        -0.412     0.723
Умный         0.44       0.404
Ябеда         -0.47      0.807
Воспитан      0.166      0.862
Хвостун       0.121      0.758
Умелый        0.303      0.481

```

Драчун	0.025	0.867
Шалун	-0.171	0.86
Веселый	0.0725	0.924

Корреляционная матрица (сокращенно):

	Верн.друг	Смелый	Красивый	Добрый	Хитрый	Жадный	Плакса
Смелый	0.482						
Красивый	-0.103	-0.218					
Добрый	0.99	0.448	-0.061				
Хитрый	-0.351	0.324	-0.18	-0.386			
Жадный	-0.843	-0.29	-0.174	-0.838	0.622		
Плакса	-0.128	-0.822	0.255	-0.14	-0.573	-0.146	
Умный	0.817	0.633	0.133	0.834	-0.237	-0.683	-0.452
Ябеда	-0.473	-0.726	-0.109	-0.461	-0.202	0.207	0.506
Воспитан	0.426	0.048	0.624	0.417	-0.605	-0.69	0.281
Хвастун	-0.405	0.004	-0.236	-0.379	0.799	0.603	-0.427
Умелый	0.699	0.579	0.287	0.699	-0.282	-0.681	-0.353
Драчун	-0.499	0.225	-0.316	-0.543	0.94	0.682	-0.449
Шалун	0.2	0.331	-0.525	0.196	0.67	0.188	-0.498
Веселый	0.64	0.659	-0.248	0.671	0.192	-0.389	-0.71

• • •

Собственные значения и процент объясняемой дисперсии факторов

Фактор:	1	2	3	4	5	6
Собств. зн	6.57	4.79	1.58	0.853	0.663	0.335
Дисперс%	43.8	32	10.6	5.69	4.42	2.23
Накоплен%	43.8	75.8	86.3	92	96.4	98.7

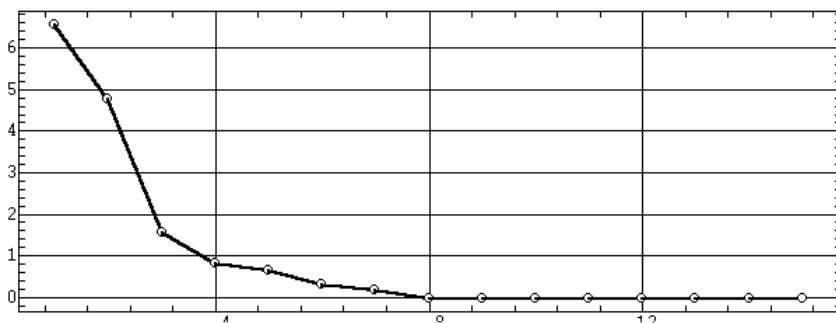


Рис. 11.12. График собственных значений факторов (по оси Y) относительно номеров факторов (по оси X)

**Обсуждение:** Как видно из результатов и рис. 11.12 в ходе вычисления главных компонент можно выделить компактную систему из трех–четырёх основных факторов, отражающих 86–92% дисперсии объектов–героев (вместо исходного 15–мерного пространства свойств).

**Продолжение результатов:**

Переменная <Собственные вектора (коэфф. поворота факторных осей)>

Верн.друг	0.334	0.141	-0.271	-0.0457	0.257	-0.0888
Смелый	0.163	0.351	0.231	-0.309	-0.139	-0.312
Красивый	0.0953	-0.198	0.496	0.525	0.4	0.129
Добрый	0.337	0.137	-0.276	0.0229	0.241	0.0947
Хитрый	-0.229	0.31	0.247	0.115	0.181	-0.299
Жадный	-0.348	0.0346	0.197	-0.0774	-0.137	0.464

Плакса	-0.0217	-0.399	-0.256	0.02	0.392	-0.219
Умный	0.347	0.173	0.0808	0.137	-0.149	0.185
Ябеда	-0.168	-0.231	-0.338	0.435	-0.51	-0.356
Воспитан	0.305	-0.209	0.255	0.166	-0.0395	-0.34
Хвастун	-0.27	0.261	-0.0646	0.455	0,0559	0.0881
Умелый	0.349	0.0905	0.225	0.141	-0.239	-0.0828
Драчун	-0.293	0.257	0.154	-0.0481	0.12	-0.458
Шалун	-0.108	0.379	-0.307	0.159	0.301	-0.00969
Веселый	0.175	0.359	-0.163	0.341	-0.208	0.123

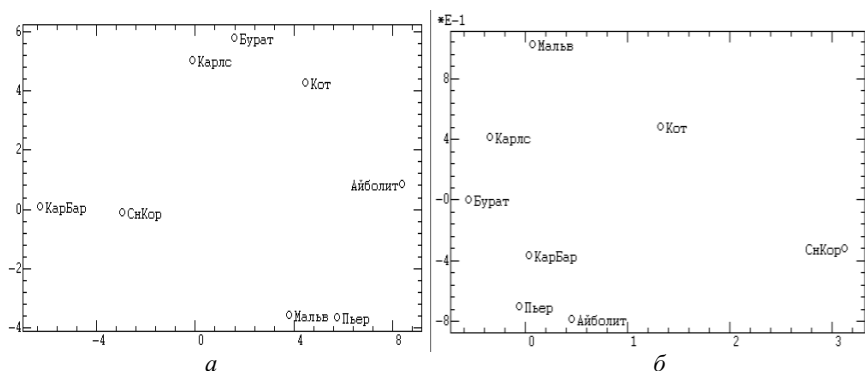


Рис. 11.13. Объекты (герои) в проекции на плоскость главных компонент:

$a$  — факторная плоскость 1–2;  $b$  — факторная плоскость 3–4

**Обсуждение:** Как видно из числовой выдачи и рис. 11.13 исследуемые объекты–герои в проекции на плоскости первых трех факторов образуют хорошо различимые группировки, что дает надежду на хорошую дальнейшую интерпретацию результатов.

### Продолжение результатов:

Переменная	<----- Факторные нагрузки до вращения ----->						
Верн. друг	0.857	0.309	-0.341	-0.0422	0.209	-0,0514	-0.0757
Смелый	0.417	0.769	0.291	-0.285	-0.113	-0.18	0.152
Красивый	0.244	-0.433	0.625	0.485	0.325	0.0745	0.127
Добрый	0.864	0.3	-0.348	0.0212	0.196	0,0548	-0.023
Хитрый	-0.588	0.679	0.311	0.106	0.147	-0.173	-0.183
Жадный	-0.892	0.0757	0.248	-0.0714	-0.112	0.268	-0.217
Плакса	-0,0555	-0.875	-0.322	0.0185	0.319	-0.127	-0.0973
Умный	0.89	0.378	0.102	0.126	-0.121	0.107	-0.11
Ябеда	-0.432	-0.506	-0.425	0.402	-0.415	-0.206	-0.0155
Воспитан	0.783	-0.456	0.32	0.153	0.0322	-0.197	-0.11
Хвастун	-0.692	0.572	-0.0813	0.42	0.0455	0,051	0.0739
Умелый	0.895	0.198	0.283	0.13	-0.194	-0.048	-0.15
Драчун	-0.751	0.563	0.194	-0.0444	0.0978	-0.265	-0.003
Шалун	-0.277	0.829	-0.387	0.147	0.245	-0.00561	-0.0616
Веселый	0.449	0.787	-0.205	0.315	-0.169	0.0712	0.0614

**Обсуждение:** Как видно из числовой выдачи и рис. 11.14 исходные свойства проецируются в сравнимой степени на несколько факторов, что может затруднить их предметную интерпретацию. Поэтому выберем три основных фактора и проведем их варимакс–вращение в пространстве 15



исходных переменных (свойств) с целью получения более простой структуры проекции.

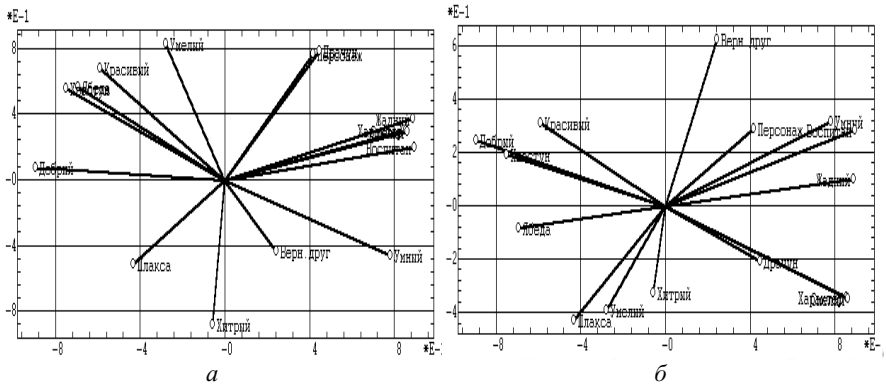


Рис. 11.14. Факторные нагрузки исходных свойств героев:  
а — факторная плоскость 1–2; б — факторная плоскость 3–4

#### Продолжение результатов:

Переменная	Общность	Специфичность
Верн. друг	0.946	0.0546
Смелый	0.85	0.15
Красивый	0.637	0.363
Добрый	0.957	0.043
Хитрый	0.904	0.0968
Жадный	0.863	0.137
Плакса	0.872	0.128
Умный	0.945	0.055
Ябеда	0.623	0.377
Воспитан	0.924	0.0759
Хвостун	0.813	0.187
Умелый	0.92	0.0799
Драчун	0.918	0.0821
Шалун	0.914	0.0863
Веселый	0.863	0.138

Переменная	<----->	Факторные нагрузки после вращения	----->
Верн. друг	0.359	-0.897	-0.114
Смелый	0.894	-0.178	-0.14
Красивый	0.0865	0.0875	0.789
Добрый	0.351	-0.907	-0.111
Хитрый	0.48	0.694	-0.438
Жадный	-0.125	0.897	-0.207
Плакса	-0.864	-0.154	0.32
Умный	0.659	-0.698	0.154
Ябеда	-0.768	0.134	-0.123
Воспитан	0.0967	-0.532	0.795
Хвостун	0.152	0.58	-0.674
Умелый	0.62	-0.618	0.393
Драчун	0.271	0.769	-0.503
Шалун	0.333	0.0791	-0.892
Веселый	0.655	-0.455	-0.476

**Обсуждение:** Как видно из сравнения рис. 11.14 и 11.15, в результате вращения получена более простая структура, в которой исходные переменные (свойства) преимущественно проецируются на один из трех главных факторов, что облегчает задачу интерпретации факторов в терминах исходных переменных.

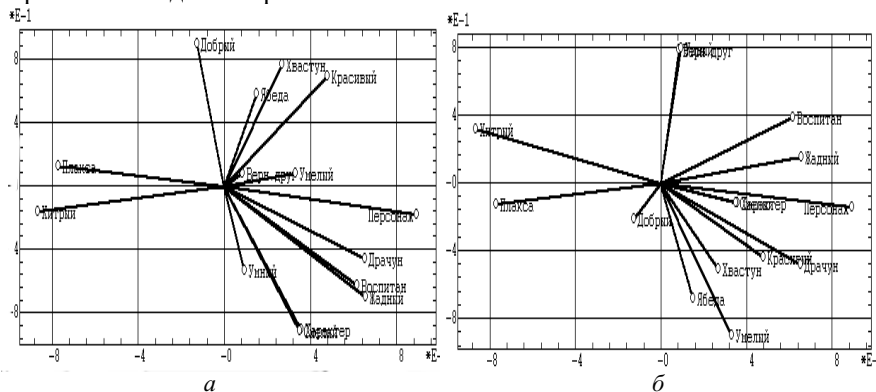


Рис. 11.15. Факторные нагрузки исходных свойств героев после вращения:  
а — факторная плоскость 1–2; б — факторная плоскость 3–4

Напомним, что факторная нагрузка показывает, насколько выражено в данной шкале (в переменной) содержание, которое описывает фактор. При интерпретации важно найти семантический смысл выделенных факторов, чему помогает группировка объектов и свойств по наибольшим нагрузкам основных факторов (см. табл. 11.2).

Таблица 11.2. Группировка свойств и героев по факторам

	Фактор 1	Фактор 2	Фактор 3
Отрицательн. Свойства	Плакса=-0.84 Ябеда=-0.768	Жадный=0.897 Драчун=0.769 Хитрый=0.694	Шалун=-0.892 Хвостун=-0.674
Отрицательн. Герои	КарВараб=-0.823 Мальвина=-0.52 Пьеро=-0.239	Кар-Вараб=2.48 СнежнКорол=1.88	Карлсон=-2.75 Буратино=-2.16 Кар-Вараб=-1.86
Положительн. Свойства	Смелый=0.894 Умный=0.659 Веселый=0.655 Умелый=0.62	Добрый =-0.907 Верн. друг=0.897 Умный=-0.698 Умелый=-0.618	Воспитан=0.795 Красивый=0.798
Положительн. герои	Кот в сап=2.18 Айболит=1.52	Айболит=-3.19 Пьеро= -2.35	Пьеро=3.49 Мальвина=2.97 Айболит=2.26

После исследования содержимого табл. 11.2 можно дать следующие психологические интерпретации трех факторов:

- фактор 1: «Сила личности (эго)»;
- фактор 2: «Добро — Зло»;
- фактор 3: «Социальная нормативность — акцентированность».

Следует обратить особое внимание на то, что категория *Сила личности* для современных взрослых воспитателей стоит на более высоком уровне, чем дихотомия *Добро—Зло*.

**Упражнение.** В качестве самостоятельного упражнения полезно заполнить исходную матрицу *герои—свойства* своими собственными ответами или ответами своего ребенка и, повторив факторный анализ, сравнить полученные результаты с нормативными. Следует, однако, предостеречь от резких педагогических выводов из такого сопоставления, поскольку приведенные нормативные результаты отражают только представления взрослого человека о добре и зле и применительно не к самому себе, а к идеальному с их точки зрения ребенку, что может быть далеко не вполне толерантно представлениям об этих категориях нормального ребенка.

**Заключение.** Если заменить в данной методике сказочных героев на политических лидеров, партии, государства или транснациональные корпорации, легко осознать значимость получаемых результатов и практическую ценность выводов в совершенно других и неизмеримо менее «игрушечных» областях.<sup>1</sup>

---

## Пример 2

Классический дидактический пример социологического применения факторного анализа приведен в статье Д.К. Арчера, Ф.М. Шелли, П.Д. Тейлора и Э.Р. Уайта<sup>2</sup> для изучения факторов, влияющих на президентские выборы в США (рис. 11.16). *Сырые* исходные данные представляют собой матрицу, содержащую проценты голосов, поданных за демократов на 28 выборах в 1872—1984 гг. (строки) в каждом из 50 штатов (столбцы).

На основе этих данных независимо решались две следующие задачи:

- 1) **выделение региональных факторов:** в этом случае строилась матрица «штат—штат» корреляций временных рядов изменения процента голосов (между строками в исходной матрице);
- 2) **выделение временных факторов:** в этом случае строилась матрица «выборы—выборы» корреляций между процентами голосов всех штатов (между столбцами в исходной матрице).

---

<sup>1</sup> Об использовании факторного анализа в социальной и политической психологии см. статьи В.Ф.Петренко с сотрудниками в Психологическом журнале: №6, 1991 (Семантическое пространство политических партий России) и №6, 1995 (Психосемантический анализ динамики качества жизни Россиян).

<sup>2</sup> Scientific American, №9, 1988.

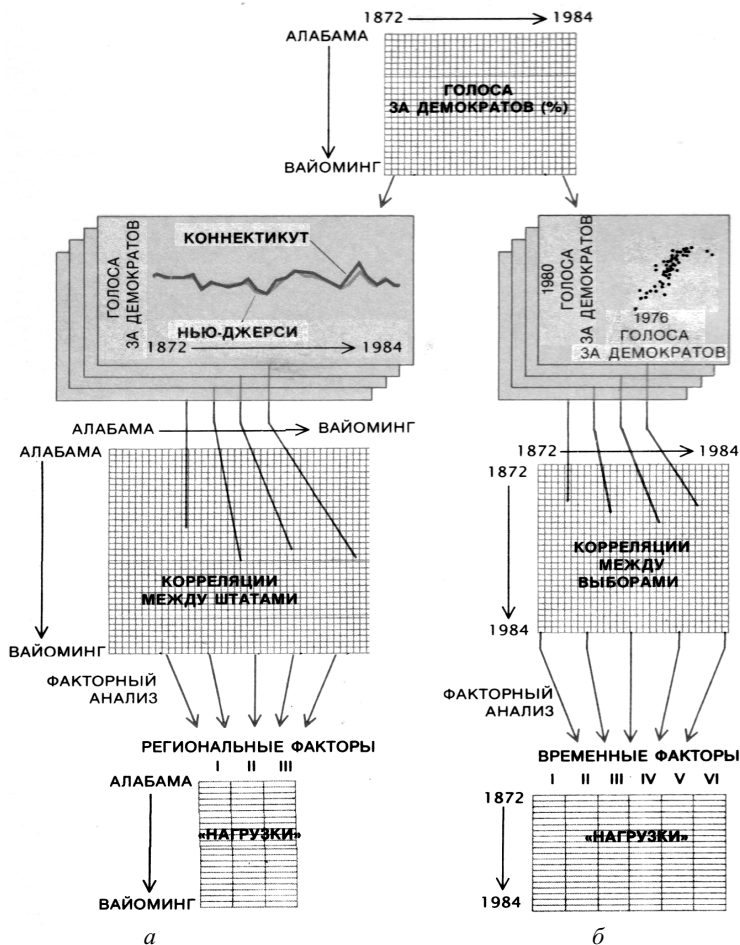


Рис. 11.16. Факторный анализ президентских выборов в США. Из исходной матрицы «число голосов, поданных за демократов» по штатам и годам выделяются:

$a$  — временные зависимости по каждому штату;  $b$  — зависимости между всеми парами выборов, по всем выборам каждая. Для этих двух типов данных вычисляются корреляционные матрицы, которые подвергаются факторному анализу. В результате выделено три основных действующих региональных и шесть временных факторов.

Эти две корреляционные матрицы и явились исходными данными для факторного анализа. Данный пример интересен еще и тем, что в нем проводятся анализ прямой и транспонированной матриц, т. е. матрицы «переменные—измерения» и матрицы «измерения—переменные».

В результате оказалось, что большинство выявленных *региональных корреляций* можно описать всего тремя факторами (из 50 штатов),

интерпретируемых как «нормальный региональный», «нормальный либеральный» и «нормальный консервативный» типы. Во *временном срезе* было выделено шесть основных факторов (из 28 выборов), интерпретируемых по двум типам «предпочтения»: 1) *социально–экономические*: «городской», «сельский», «пригородный»; 2) *региональные*: «юг», «запад», «северо–восток».

Каждому штату соответствует определенная *нагрузка* по каждому фактору (проекция фактора на штат) и для каждых выборов. И здесь мы имеем также прекрасный пример использования графиков изменения факторных нагрузок во времени.

Так в результате изучения временных графиков изменения нагрузок региональных факторов было выявлено существование «избирательных эпох», в каждой из которых в отдельных штатах проявляется тенденция сохранять определенный тип голосования по всей стране (рис. 11.17).

При изучении графиков изменения нагрузок временных факторов было выявлено различие региональных предпочтений перед социально–экономическими и устойчиво–периодический характер изменения обоих типов предпочтения (рис. 11.18).

Добавим к этому, что подобные графики являются плодотворным материалом для последующего построения прогностических моделей методами регрессионного и фурье–анализа.

В результате дальнейшего осмысления результатов были сформулированы и верифицированы многие важные комплексные понятия, например: *урбанизация и региональное разделение, устойчивые и неустойчивые предпочтения* и др.

Чрезвычайно эффективными для визуального исследования явились также географические карты, раскрашиваемые по различным показателям факторного анализа (рис. 11.19).

Таким образом, визуальное изучение наглядных факторных (географических) диаграмм и их временных графиков позволяет понять динамику и механизмы сложных социальных процессов и дать им практическую интерпретацию, важную для выработки текущей экономической и региональной политики, а также для коррекции последующих предвыборных стратегий.

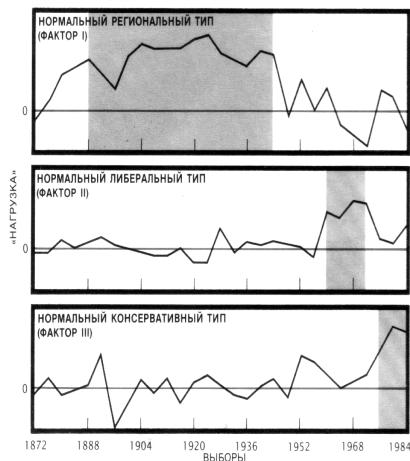


Рис. 11.17. «Избирательные эпохи», в течение которых поведение избирателей в конкретном регионе остается достаточно стабильным

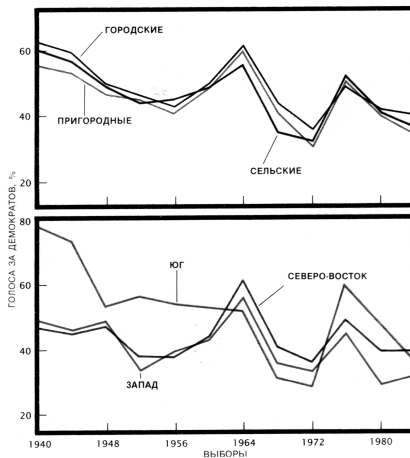


Рис. 11.18. Региональные предпочтения (внизу) играют большую роль (большее различие нагрузок), чем социально-экономические (вверху)

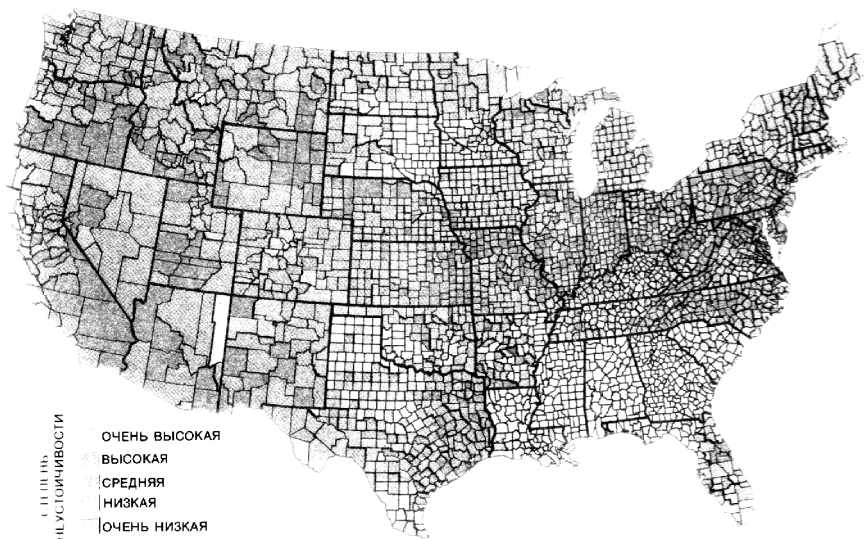


Рис. 11.19. Неустойчивость предпочтений избирателей, равно как и стабильность их симпатий, географически остается неизменной (данные по более 3000 графств).

Районы с наибольшей неустойчивостью (Северо-Восток, Юг, север Среднего Запада) составляют главные поля сражения обеих партий

**Заключение.** Заменим в данном примере процент голосов любым экономическим показателем и получим множество впечатляющих

примеров факторных исследований из области маркетинга, менеджмента, бизнеса, экономики, геополитики и т. п.

### Пример 3

**З а д а ч а.** При приеме на работу 18 претендентов прошли 10 специальных тестов, результаты каждого из которых оценивались по 10–балльной системе (табл. 11.3, файл TESTS<sup>1</sup>). Отметим, что этот тип данных в аналитическом плане эквивалентен многочисленным задачам различного типа опросов с ответами в заданной шкале предпочтений.

Интересно было бы выявить главные факторы, действующие в этой батарее из 10 тестов, а также выяснить наличие или отсутствие группировок среди претендентов.

Таблица 11.3. Результаты тестирования 18 претендентов на работу по 10–балльной системе

Тесты: — Испы- туемые	Память на числа	Мате- матиче- ские задачи	Наход- чивость в диалог е	Состав- ление алго- ритмов	Уверен- ность в выступ- лении	Коман- дный дух	Наход- чи- вость	Сотруд- ничест- во	Призна- ние в коллек- тиве	Сила убежде- ния
1	10	10	9	10	10	10	9	10	10	9
2	10	10	4	10	5	5	4	5	4	3
3	5	4	10	5	10	4	10	5	3	10
4	10	10	9	10	10	10	9	10	10	9
5	4	3	5	4	3	10	4	10	10	5
6	10	10	4	10	5	4	3	4	5	5
7	4	4	5	5	4	10	5	10	10	6
8	4	5	3	4	5	10	4	10	10	4
9	4	5	10	4	10	5	10	4	3	10
10	10	10	4	10	5	4	4	5	4	4
11	4	5	10	5	10	4	10	4	5	10
12	10	10	9	10	10	9	9	10	10	10
13	6	5	4	3	5	10	5	10	10	5
14	4	5	10	4	10	5	10	3	4	10
15	10	10	9	10	10	9	10	9	10	10
16	6	5	3	4	4	10	4	10	10	5
17	10	10	5	10	4	5	4	3	4	5
18	4	5	10	4	10	4	10	4	4	10

Здесь мы имеем дело с ранговыми переменными, поэтому для оценки их коррелированности приемлемы коэффициенты Спирмана, Кенделла и Крамера. Для сохранения общности результатов с номинальными

<sup>1</sup> Данные из архива SPSS, аналитика наша.

данными используем в факторном анализе коэффициент связности Крамера.

### Результаты:

ФАКТОРНЫЙ АНАЛИЗ. Файл: tests.std

Собственные значения и процент объясняемой дисперсии факторов

Фактор:	1	2	3	4	5	6	7	8
Собств.зн	5.72	1.35	1.02	0.52	0.423	0.327	0.269	0.199
Дисперс%	57.2	13.5	10.2	5.2	4.23	3.27	2.69	1.99
Накоплен%	57.2	70.6	80.9	86.1	90.3	93.6	96.2	98.2

Переменная <---- Факторные нагрузки до вращения ---->

Фактор:	1	2	3	47	8
память	0.752				
мат.зада	0.75				
прям.диа	0.885				
алгоритм	0.736				
уверенно	0.717		-0.558		
команд.д	0.715	0.538			
находчив	0.83				
сотрудни	0.724	0.538			
признани	0.697	0.517			
убеждени	0.735				

### Выводы.

1. Как видно из таблицы собственных значений, в качестве главных факторов можно с уверенностью выбрать первые три, которые покрывают 80,9% дисперсии измерений. Четвертый и последующие компоненты в 2 с лишним раза меньше по величине третьего компонента, к тому же их собственные значения существенно меньше единицы.

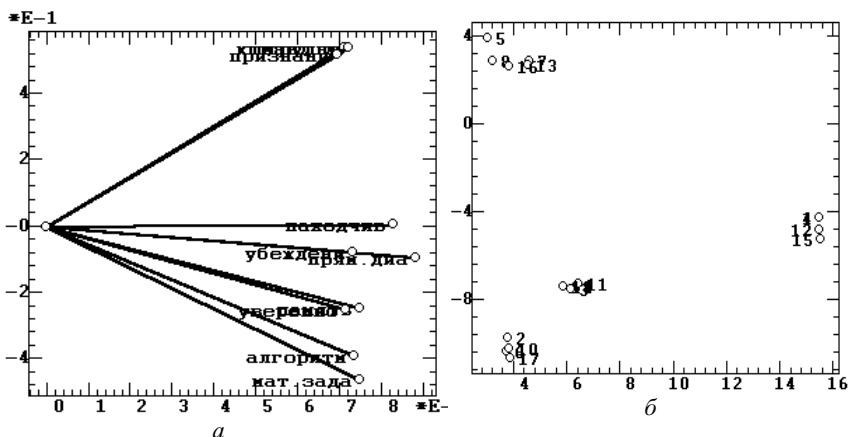


Рис. 11.20. Проекция на плоскость факторов 1–2 результатов тестирования: а — факторные нагрузки; б — претенденты

2. Как видно из таблицы факторных нагрузок, все тесты преимущественно проецируются на первый фактор, который можно было бы назвать фактором «общей профессиональной пригодности».



На второй фактор проецируются с недоминирующими весами тесты *командный дух, сотрудничество, признание в коллективе*. Этот фактор можно было бы назвать «эффективность в коллективной работе». На третий фактор с недоминирующим весом проецируется всего лишь один и малозначимый на начальном этапе работы претендента тест *уверенность в выступлении*. В связи с этим третий фактор можно в предварительном анализе не рассматривать.

3. Как видно из графиков факторных нагрузок (рис. 11.20, а), две пары тестов имеют очень близкие по величине и проекции нагрузки: *командный дух, сотрудничество* и *уверенность в выступлении, память*. Поэтому эти тесты при повторном факторном анализе можно было бы исключить из рассмотрения.
4. В проекции претендентов (рис. 11.20, б) выделяются четыре плотные и далеко удаленные друг от друга группы претендентов: 1)={5, 7, 8, 13, 16}; 2)={2, 6, 10, 17}; 3)={3, 9, 11, 14, 18}; 4)={1, 4, 12, 15}. Группа 4 имеет заметное преимущество по первому фактору, а группа 1 — по второму фактору. Группы же 2 и 3 показали посредственную пригодность по обоим факторам, поэтому по их представителям, вероятно, следует вынести отрицательное решение в отношении приема на работу.

#### *Последующие исследования.*

1. Провести факторный анализ с использованием корреляций Спирмана и Кенделла и оценить различие или сходство в результатах.
2. Исходные данные можно преобразовать к бинарному виду, например, по следующим соображениям. Установить порог оценок: удовлетворительно—неудовлетворительно (скажем, по уровню 5 баллов) и преобразовать по этому порогу данные<sup>1</sup>. Затем провести факторный анализ с использованием различных формул бинарных коэффициентов и оценить различие или сходство в результатах.
3. Провести кластерный анализ по аггломеративным и дивизивной стратегиям и сравнить выявленные кластеры с группировками, визуально выявленными на рис. 11.20, б.

Эти задачи мы оставляем в качестве учебной практики читателям.

## **11.2. Кластерный анализ**

### *Пример 1*

**Задача.** Произведено измерение четырех важных показателей для 20 популярных сортов немецкого пива (табл. 11.4, файл CLA).

*Таблица 11.4. Основные показатели популярных сортов немецкого пива*

<sup>1</sup> Операцией кодирования в Блоке преобразований по условиям  $x > 5$  и  $x < 6$ .

№	Сорт	Калории	Натрий	Алкоголь	Цена
1	Budweiser	144	15	4.7	0.43
2	Schlitz	151	19	4.9	0.43
3	Lowenbraw	157	15	4.9	0.48
4	Kronenbourg	170	7	5.2	0.73
5	Heineken	152	11	5	0.77
6	Old_Milwaukee	145	23	4.6	0.28
7	Augsberger	175	24	5.5	0.4
8	Strohs_Bonemain_Style	149	27	4.7	0.42
9	Miller_Lighr	99	10	4.3	0.43
10	Bodweiser_Light	113	8	3.7	0.44
11	Coors	140	18	4.6	0.44
12	Coors_Light	102	15	4.1	0.46
13	Mihelob_Light	135	11	4.2	0.5
14	Becks	150	19	4.7	0.76
15	Kirin	149	6	5	0.79
16	Pabst_Extra_Light	68	15	2.3	0.38
17	Hamms	136	19	4.4	0.43
18	Heilemans_Old_Stile	144	24	4.9	0.43
19	Olimpia_Gold_Light	72	6	2.9	0.46
20	Schlitz_Light	97	7	4.2	0.47

Необходимо произвести кластеризацию этих данных с использованием евклидовой метрики для выявления имеющихся группировок. Это поможет рекламировать и потреблять более дешевые сорта пива, но близкие по показателям к более дорогим.

### Результаты:

КЛАСТЕРНЫЙ АНАЛИЗ. Файл: cla.std

Эвклид+Ближ.сосед

Таблица расстояний

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
( 2)	8.06						
( 3)	13	7.21					
( 4)	27.2	22.5	15.3				
( 5)	8.96	8.07	6.41	18.4			
( 6)	8.06	7.22	14.4	29.7	13.9		
( 7)	32.3	24.5	20.1	17.7	26.4	30	
( 8)	13	8.25	14.4	29	16.3	5.66	26.2
( 9)	45.3	52.8	58.2	71.1	53	47.8	77.3

• • •

К л а с т е р ы : (список объектов) -> расстояние

```

(14,2) --> 1.07
(18,6) --> 1.45
(20,9) --> 3.61
(17,11) --> 4.13
(17,1,11) --> 5
(18,8,6) --> 5.66
(15,5) --> 5.83
(20,12,9) --> 5.83
(15,3,5) --> 6.41
(18,14,2,8,6) --> 6.42
(18,17,1,11,14,2,8,6) --> 7.07
(18,15,3,5,17,1,11,14,2,8,6) --> 7.21
(18,13,15,3,5,17,1,11,14,2,8,6) --> 8.07

```

(19,16) --> 9.87  
 (20,10,12,9) --> 13  
 (18,4,13,15,3,5,17,1,11,14,2,8,6) --> 15.3  
 (18,7,4,13,15,3,5,17,1,11,14,2,8,6) --> 17.7  
 (20,18,7,4,13,15,3,5,17,1,11,14,2,8,6,10,12,9) --> 22.2  
 (20,19,16,18,7,4,13,15,3,5,17,1,11,14,2,8,6,10,12,9) --> 25.1

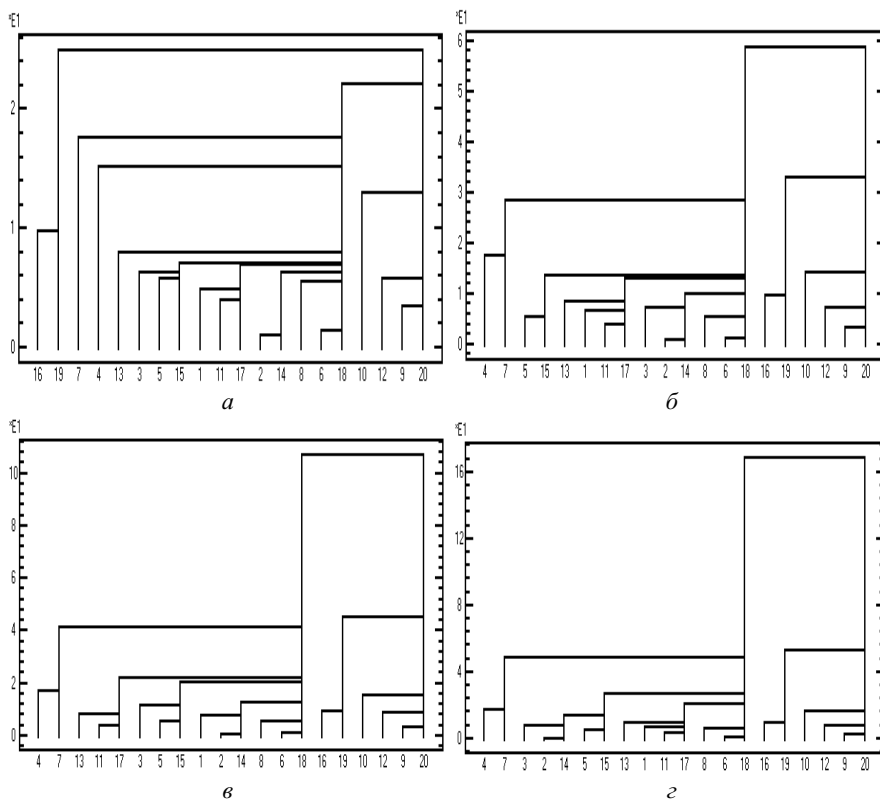


Рис. 11.25. Дендрограммы (по оси  $Y$  — расстояние объединения, по оси  $X$  — номера объектов):  
 а — стратегия ближайшего соседа; б — стратегия группового соседа; в — стратегия дальнего соседа; г — стратегия Урда

**Обсуждение результатов:** Как видно из рис. 11.25, а, стратегия ближайшего соседа достаточно отчетливо выделяет три группы сортов пива: (1–3, 5, 6, 8, 11, 13–15, 17, 18); (9, 10, 12, 20); (16, 19); при этом сорта 7 и 4 несколько выпадают из этой классификации. Поэтому можно попытаться применить другую, растягивающую пространство, стратегию для более яркого разделения кластеров.

Для сравнения на рис. 11.25, б–г приведены дендрограммы, построенные с использованием стратегии группового и дальнего соседа и стратегии Урда.

Как можно заметить из сравнения дендрограмм, стратегии группового соседа, дальнего соседа и Уорда по сравнению со стратегией ближайшего соседа дают все более четкое выделение кластеров и несколько отличные классификации. Так на дендрограммах рис. 11.25, *в, г* можно выделить 3–5 кластеров. В связи с этим далее применим дивизивную стратегию с эвклидовой метрикой, чтобы сравнить группировки исследуемых объектов на различное число кластеров.

### Результаты:

КЛАСТЕРНЫЙ АНАЛИЗ. Файл: cla.std

Эвклид+Дивизивная

К л а с т е р ы:

Среднее внутрикластерное расстояние=9.81

- 1= (Miller Li\*, Bodweiser, Coors Lig, Pabst Ext, Olimpia G, Schlitz L)
- 2= (Schlitz, Lowenbraw\*, Kronenbou, Heineken, Augsberge, Becks, Kirin)
- 3= (Budweiser, Old Milwa, Strohs Bo, Coors\*, Mihelob L, Hamms, Heilemans)

К л а с т е р ы:

Среднее внутрикластерное расстояние=7.37

- 1= (Pabst Ext\*, Olimpia G)
- 2= (Schlitz, Lowenbraw\*, Kronenbou, Heineken, Augsberge, Becks, Kirin)
- 3= (Budweiser, Old Milwa, Strohs Bo, Coors\*, Mihelob L, Hamms, Heilemans)
- 4= (Miller Li\*, Bodweiser, Coors Lig, Schlitz L)

К л а с т е р ы:

Среднее внутрикластерное расстояние=6.4

- 1= (Pabst Ext\*, Olimpia G)
- 2= (Schlitz, Old Milwa\*, Strohs Bo, Becks, Heilemans)
- 3= (Budweiser\*, Lowenbraw, Heineken, Coors, Mihelob L, Kirin, Hamms)
- 4= (Miller Li\*, Bodweiser, Coors Lig, Schlitz L)
- 5= (Kronenbou\*, Augsberge)

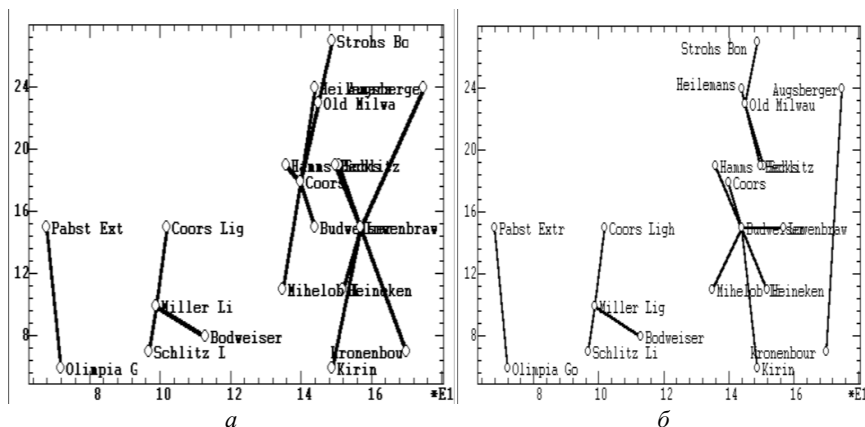


Рис. 11.26. Дивизивная кластеризация сортов пива в проекции на плоскость калории (ось *X*)—натрий (ось *Y*):  
*а* — на 4 кластера; *б* — на 5 кластеров

**Обсуждение:** Как показывают числовые результаты, переход от трех к четырем кластерам влечет существенное повышение компактности группировки, о чем свидетельствует значительное уменьшение среднего

внутрикластерного расстояния (с 9,81 до 7,37). Переход же от четырех к пяти кластерам сопровождается более чем в 2 раза меньшим сокращением внутрикластерного расстояния (с 7,37 до 6,4). Тем самым группировку на три кластера следует признать неудачной из-за большого внутрикластерного расстояния.

Чтобы выбрать лучшую группировку из оставшихся двух, близких по внутрикластерному расстоянию, обратимся к визуальному анализу соответствующих им диаграмм кластеризации (рис. 11.26). Из сравнения диаграмм видно, что разбиение на пять кластеров (рис. 11.26, б) менее логично, поскольку приводит к выделению кластера из двух, очень далеко отстоящих друг от друга сортов Kronenbou и Augsberge.

Из визуального исследования диаграмм можно также вывести еще одно интересное наблюдение: кластеризация сортов пива идет преимущественно по оси калорий. В чем здесь дело? Обратившись к нашим исходным данным (табл. 11.4), можно заметить существенную разницу в значениях переменных: значение переменной *калории* в 10 раз превышает значения переменной *натрий* и еще больше превышают значения других переменных. Поэтому и удаленность объектов по шкале калорий подавляюще превалирует при кластеризации по сравнению с удаленностью по другим шкалам. Чтобы сделать переменные равноправными, следует вместо евклидовой метрики использовать нормированную евклидову метрику.

### Результаты:

```

КЛАСТЕРНЫЙ АНАЛИЗ.  Файл: cla.std  Норм.Евклид.+Дивизивная
                        К л а с т е р ы:
Среднее внутрикластерное расстояние=0.813
1= (Pabst Ext*,Olimpia G)
2= (Kronenbou,Heineken*,Becks,Kirin)
3= (Budweiser,Schlitz*,Lowenbraw,Old Milwa,Augsberge,Strohs
   Bo,Coors,Hamms,Heilemans)
4= (Miller Li*,Bodweiser,Coors Lig,Mihelob L,Schlitz L)

```

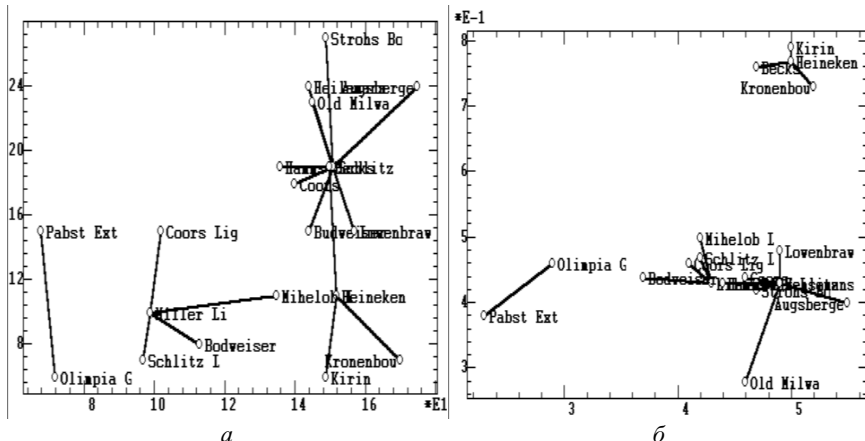


Рис. 11.27. Результаты дивизивной кластеризации в нормированной евклидовой метрике:

*а* — плоскость калорий—натрий; *б* — плоскость алкоголь—цена

**Обсуждение:** Как видно из диаграмм кластеризации (рис. 11.27) теперь группировка сортов пива идет не только по шкале калорий, но и по трем другим шкалам, тем самым более адекватно учитывая вклады всех переменных.

Сравнивая две диаграммы рис. 11.26, *а* и 11.27, *а*, следует отметить малое изменение двух кластеров в области небольших калорийностей и заметную перегруппировку двух кластеров в области высоких калорийностей.

Далее можно задаться вопросом, а все ли переменные одинаково значимы для кластеризации, нет ли среди них второстепенных, которые можно было бы не учитывать. Например, могут быть пары переменных, связанных четкой функциональной зависимостью, тогда из такой пары можно в анализе учитывать только одну переменную, поскольку она однозначно определяет значения другой переменной. Для поиска таких связанных переменных можно было бы воспользоваться корреляционным анализом, выбрать пары с высокими коэффициентами корреляции, а затем рассмотреть их диаграммы рассеяния для визуальной оценки степени функциональной зависимости. Однако можно поступить проще — провести еще одну кластеризацию, но уже не объектов, а переменных. Для этого надо просто использовать метрику коэффициентов корреляции. При этом выберем стратегию ближайшего соседа, сжимающую пространство, чтобы на дендрограмме визуально выделялись только очень близкие друг к другу группировки.

### Результаты:

КЛАСТЕРНЫЙ АНАЛИЗ. Файл: cla.std Корреляция+Ближ.сосед  
К л а с т е р ы:

(список объектов) -> расстояние  
 (3,1) --> 0.079  
 (3,2,1) --> 0.588  
 (4,3,2,1) --> 0.668  
 1=калории, 2=натрий, 3=алкоголь, 4=цена

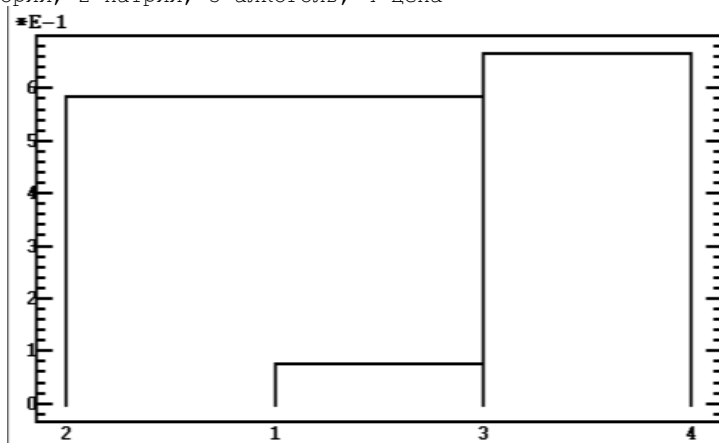


Рис. 11.28. Дендрограмма классификации переменных: по оси Y — расстояние, по оси X — номер переменной

**Обсуждение:** Как видно из числовой выдачи и дендрограммы (рис. 11.28), две переменные 1 и 3 (содержание калорий и алкоголя) очень близки друг к другу (коррелированы): они объединены на расстоянии 0,079, тогда как следующие два объединения происходят на расстояниях в 7 с лишним раз больших. Поэтому одну из этих переменных можно было бы вполне исключить из рассмотрения, повторить кластерный анализ и сравнить результаты.

Интересно было бы также провести факторный анализ, выяснить, какие факторы являются главными, вычислить координаты объектов в пространстве главных факторов и уже в этом пространстве провести новую кластеризацию и сравнить полученные результаты.

Предлагаем эти задачи в качестве учебной практики читателям. Дополнительные примеры разнопланового применения кластерного анализа рассмотрены в гл. 14. Анализ данного же примера будет продолжен в следующем разделе.

## Пример 2

**Задача.** Возвратимся к данным тестирования профессиональной пригодности примера 3 к разд. 11.1 и к результатам их факторного анализа. В частности, там были обнаружены практически совпадающие по главным факторам две пары переменных и четыре плотные группы

тестируемых. Интересно было бы методом кластерного анализа проверить эти результаты. Используем метрику коэффициента связности Крамера и стратегию дальнего соседа (для более четкого разделения кластеров).

### Результаты:

```

КЛАСТЕРНЫЙ АНАЛИЗ.  Файл: tests.std
                      Крамера+Дальн.сосед
                      К л а с т е р ы:
                      (список переменных) -> расстояние
(5,3) --> 0.216   увереннос, прям. диал
(8,6) --> 0.225   сотруднич, команд. ду
(4,1) --> 0.247   алгоритмы, память
(4,2,1) --> 0.29  алгоритмы, мат. задач, память
(10,7) --> 0.322  убеждение, находчиво
(9,8,6) --> 0.358 признание, сотруднич, команд. ду
(10,5,3,7) --> 0.379 убеждение, увереннос, прям. диал, находчиво
(10,4,2,1,5,3,7) --> 0.664
убеждение, алгоритмы, мат. задач, память, увереннос, прям. диал, находчиво
(10,9,8,6,4,2,1,5,3,7) --> 0.718

```

**В ы в о д ы:** Как видно из числовой выдачи, кластерный анализ подтвердил близость только одной пары тестов: *командный дух* и *сотрудничество*, которые были объединены на втором шаге кластеризации с расстоянием 0,225. Что касается теста *уверенность в выступлении*, то он объединен не с тестом на *память*, а с *находчивостью в прямом диалоге*, тест же *память* оказался близок к *составлению алгоритмов*.

Такие изменения становятся понятными, если мы возвратимся к графикам факторных нагрузок (рис. 11.20, а), на которых видно, что пара тестов *уверенность в выступлении* — *память*, непосредственно соседствует именно с *находчивостью в прямом диалоге* и *составлением алгоритмов*. Смещение связей объясняется тем, что кластерный анализ использует всю информацию о тестах, в то время как проекция нагрузок на плоскость первых двух факторов (рис. 11.20, а), как это следует из процента накопленной дисперсии в таблице факторных нагрузок, отражает только 70,6% исходной информации, поэтому в проекциях на другие факторные плоскости эти тесты могут сочетаться в других комбинациях.



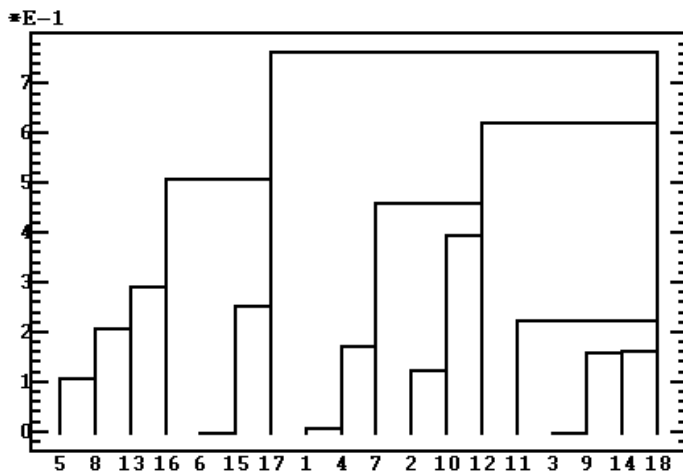


Рис. 11.29. Дендрограмма классификации претендентов

*Продолжение анализа.* Теперь можно перейти к классификации претендентов с использованием той же метрики Крамера. Такое применение метрики допустимо, поскольку переменные однородны (все являются тестами) и число градаций их значений одинаково. Перед повторением кластерного анализа надо предварительно транспонировать матрицу данных, чтобы претенденты располагались по столбцам, а тесты — по строкам (поскольку, как было отмечено выше, метрика Крамера по умолчанию классифицирует переменные, а не объекты).

**В ы в о д ы:** На полученной дендрограмме (рис. 11.29) можно выделить три–четыре кластера. Выпишем эти кластеры и сравним с визуально выделенными группами в факторном анализе (рис. 11.20, б)

КлАн: {5, 8, 13, 16}      {6, 15, 17}      {3, 9, 11, 14, 18}      {1, 2, 4, 7, 10, 12}  
 ФаАн: {5, 7, 8, 13, 16}      {2, 6, 10, 17}      {3, 9, 11, 14, 18}      {1, 4, 12, 15}

Как можно заметить, четыре группы претендентов в значительной степени сохранили свой состав. Перемещение отдельных претендентов (между первой, второй и четвертой группами, они отмечены жирным шрифтом) можно также объяснить полным использованием в кластерном анализе исходной информации.

Здесь особо следует подчеркнуть, что упомянутое полное использование информации может иметь и свои негативные стороны, поскольку вся (нерафинированная) информация включает и случайные (шумовые) компоненты, которые могут исказить или смещать главные закономерности.

*Последующие исследования.*

1. Провести кластерный анализ с использованием дивизивной стратегии с разбиением на четыре кластера и сравнить результаты.
2. Провести кластерный анализ с использованием корреляций Спирмана и Кенделла и оценить различие или сходство в результатах.
3. Исходные данные преобразовать к бинарному виду (аналогично примеру 3 к разд. 11.1). Провести кластерный анализ с использованием различных формул бинарных коэффициентов и оценить различие или сходство в результатах.

Оставляем эти задачи в качестве учебных читателям.

## 11.3. Дискриминантный анализ

### Пример 1

**З а д а ч а.** Проведем дискриминантную верификацию результатов дивизивной классификации сортов немецкого пива из примера к разд. 11.2. Для приведения исходных данных к требуемому в данном случае виду при выполнении кластерного анализа подтвердим необходимость сохранения номеров кластеров в электронной таблице. Начнем с четырех кластеров, выделенных по евклидовой метрике (рис. 11.26, а).

### Результаты:

ДИСКРИМИНАНТНЫЙ АНАЛИЗ. Файл: cla.std

Расстояние Махаланобиса=84.9, значимость=1.28E-5

Гипотеза 1: <Межкластерное расстояние отлично от нуля>

Класс <-Коэффициенты дискриминантной функции: b[0],b[1],...->

	1	-100	1.06	0.934	28.7	99.9
	2	-383	2.5	1.21	49.3	169
	3	-299	2.17	1.2	45.3	144
	4	-200	1.42	0.848	47.3	117
Объект	Класс	D^2	Значим	Апостеор.вероят.		
Budweiser	3	1.2	0.878	0.991		
Schlitz	3!	2.31	0.678	0.821		
Lowenbrow	2	2.79	0.593	0.853		
Kronenbou	2	6.23	0.182	1		
Heineken	2	1.82	0.768	1		
Old Milwa	3	1.48	0.83	1		
Augsberge	2	7.12	0.13	1		
Strohs Bo	3	3.53	0.474	0.962		
Miller Li	4	1.89	0.756	1		
Bodweiser	4	7.53	0.11	1		
Coors	3	0.143	0.998	0.998		
Coors Lig	4	1.42	0.84	1		
Mihelob L	3	3.91	0.418	1		
Becks	2	6.6	0.158	0.997		
Kirin	2	3.2	0.525	0.999		
Pabst Ext	1	2.34	0.673	1		
Hamms	3	0.772	0.942	1		
Heilemans	3	2.61	0.626	0.979		
Olimpia G	1	2.34	0.673	1		
Schlitz L	4	1.7	0.791	1		

**Обсуждение результатов:** Как показала проверка гипотезы о равенстве нулю расстояния Махаланобиса « $D^2=0$ » (ее уровень значимости очень низок — 0,0000128), кластеры достаточно компактны и хорошо разделены. Для большинства сортов пива гипотеза о нулевых расстояниях до центров соответствующих кластеров « $D_i^2=0$ » принимается с очень высокими уровнями значимости (от 0,417 до 0,998). Можно отметить сравнительно меньшие значимости для Kronenbou, Augsberge, Becks, что можно объяснить их принадлежностью ко второму достаточно разбросанному классу (рис. 11.26, а, правый кластер в области высоких калорийностей).

Обращает также на себя внимание изменение класса пива Schlitz с второго на третий. Судя по диаграмме рис. 11.26, а, этот сорт действительно находится на краю кластера 2 в непосредственной близости от граничных объектов кластера 3. Здесь проявилось определенное различие в алгоритмах кластерного и дискриминантного анализа. Дивизивные группировки кластерного анализа базируются на выделении центрального объекта в каждом кластере и поиске близлежащих к нему объектов. При этом положение такого объекта в общем не совпадает с геометрическим центром кластера. Дискриминантный же анализ оценивает группировки по максимуму дискриминирующей функции, положение которого в ряде случаев более соответствует геометрическому центру кластера.

В качестве учебной задачи оставляем читателям дискриминантный анализ результатов кластеризации в нормированной евклидовой метрике (рис. 11.27) и сравнения результатов с вышерассмотренными. Отметим только что в этом случае смен кластеров наблюдаться не будет.

## Пример 2

Таблица 11.5. Метрические данные 45 птиц

№	Крыло	Голова	Клюв	Лапы	Вес	Пол
1	25.8	9.5	4.2	14.6	380	0
2	25.1	9.6	3.8	14.5	355	0
3	24.8	9.4	3.8	15	355	1
4	25.9	9.3	3.9	15.5	375	0
5	25.8	9.6	4.5	15.6	360	2
6	24.8	9.4	4.4	14.7	355	1
7	25.6	9.6	4.1	14.5	356	0
8	25.3	9.4	4.4	15.4	360	2
9	25.5	9.9	4	15.1	330	2
10	24.2	8.6	3.9	13.6	350	1
11	25	9.3	3.9	15.1	338	2
12	24.7	9.3	3.8	14.5	335	0
13	24.3	9.7	4	13.9	310	2
14	25.1	9.3	3.9	15.2	332	2
15	24.7	8.9	3.8	15.1	340	2
16	25.5	9.4	4.1	15.2	333	0

17	25.7	9.7	3.9	13.8	323	2
18	25.9	9.7	4	15.6	325	2
19	25.1	9.4	4.6	14.9	325	0
20	24.7	8.6	3.9	13.9	345	1
21	25.3	8.8	3.9	14.3	340	1
22	26.1	9.3	3.8	15	327	1
23	26	9.4	3.9	14.3	316	0
24	24.8	8.4	4	14.4	336	1
25	25.6	8.8	4.2	15	330	2
26	24.6	8.4	4	14.8	330	2
27	25.1	8.6	3.8	14.4	329	0
28	25.1	8.3	3.9	14.5	338	1
29	24.9	8.9	3.8	14.6	314	0
30	25.3	8.4	3.8	13.9	335	1
31	24.7	9.3	4	15.4	285	0
32	24.7	8.2	3.7	14.3	320	1
33	25.4	8.6	4	13.8	310	1
34	25.7	8.6	3.9	13.8	315	2
35	23.8	8.4	3.7	14.1	300	1
36	24.8	8.8	3.9	14.3	290	0
37	25.3	8.4	4.1	14.3	305	1
38	24.9	8.6	3.8	14	292	1
39	25.9	8.3	3.8	13.8	317	0
40	25	8.2	3.8	13.3	305	1
41	26	8.5	3.7	13.7	300	2
42	25	8.4	3.8	13.3	290	1
43	25.7	8.4	3.9	14.8	300	1
44	24.8	8	3.6	13.2	294	0
45	25.3	8.2	3.6	13.3	292	1

**З а д а ч а.** В исследовании одного вида птиц фиксировались их метрические данные и половая принадлежность: 1 — самец, 2 — самка (табл. 11.5, файл BIRDS<sup>1</sup>). Последний признак у ряда особей по разным причинам не удалось определить (значение 0 в столбце «Пол» табл. 11.5). Поэтому встала задача найти классифицирующую функцию, по которой можно было бы приписать пол неопределенным особям.

**Результаты** (сокращенно):

ДИСКРИМИНАНТНЫЙ АНАЛИЗ. Файл: birds.std

Расстояние Махаланобиса=2.46, значимость=0.783

Гипотеза 0: <Межкластерное расстояние не отлично от нуля>

Класс <- Коэффициенты дискриминантной функции: a[0], a[1], ... ->

1	-1.37E3	90.2	10.6	17.1	6.48	0.732
2	-1.42E3	90.8	13.4	18.2	7.56	0.702

Объект	Класс	D^2	Значим	Вероят.отнесения
1	2*	12.9	0.024	0.699
2	2*	7.64	0.177	0.718
3	2!	5.98	0.308	0.68
4	2*	12.7	0.0263	0.752
6	2!	6.34	0.274	0.754
7	2*	5.1	0.403	0.82
12	2*	3.29	0.656	0.618
13	2	11.2	0.0467	0.848
15	1!	4.61	0.465	0.537
16	2*	0.672	0.984	0.912
19	2*	14.8	0.0111	0.932
22	2!	4.6	0.466	0.88

<sup>1</sup> Данные с сокращениями из архива SPSS.

---

23	2*	3.36	0.645	0.871
26	1!	3.24	0.662	0.801
27	1*	0.858	0.973	0.769
29	2*	2.21	0.82	0.549
31	2*	18.6	0.00228	0.96
34	1!	2.39	0.793	0.73
36	2*	6.31	0.277	0.592
39	1*	4.33	0.503	0.87
41	1!	4.94	0.423	0.732
44	1*	3.31	0.653	0.971

---

**Обсуждение результатов:** Как показала проверка гипотезы о равенстве нулю расстояния Махаланобиса (ее уровень значимости очень высок), кластеры или некомпактны, или же плохо разделены. Для окончательного вывода следует обратиться к анализу значимостей нулевых гипотез для объектов.

Из изучения сокращенной таблицы результатов можно сделать следующие выводы:

- 1) семь из 31 ранее классифицированных объектов (№ 3, 6, 15, 22, 26, 34, 41) были отнесены к противоположенному классу с высокими уровнями значимости, а для объекта №13 значимость отнесения к классу 2 оказалась ниже критической;
- 2) для четырех из 14 вновь классифицированных объектов (№ 1, 4, 19, 31) значимость отнесения к соответствующему классу оказалась также ниже критической.

В процентном отношении отмеченные отклонения не слишком велики и не могут являться достаточным основанием для признания всей классификации неудачной. Однако они могут служить основанием для более глубокого изучения ситуации. Во–первых, следует посмотреть, нет ли погрешностей в самой методике сбора исходных данных, т. е. метрологических или субъективных погрешностей (это за пределами наших возможностей, поскольку сборщики данных недоступны). Во–вторых, полезно визуально изучить пространственное распределение измерений, чтобы сделать вывод о хорошем или плохом их разделении по половому признаку. Для этого следует провести факторный анализ с выдачей диаграммы рассеяния объектов на плоскость первых двух главных компонент. Оставляем эту задачу в качестве учебной читателям.

## 11.4. Шкалирование

### Пример 1

**Задача.** В сентябре 1990 г. был проведен опрос по оценке в трехбалльной шкале взаимных различий между восемью ведущими в то время политическими деятелями (табл. 11.6, файл MSC).

Таблица 11.6. Различия между популярными политическими лидерами 1990 г.

Popov	Gidaspov	Gorbi	Bush	Eltsin	Lansberg	Ligachev	Sobchak
0	–	–	–	–	–	–	–
2	0	–	–	–	–	–	–
2	1	0	–	–	–	–	–
1	2	1	0	–	–	–	–
1	2	2	2	0	–	–	–
1	3	2	1	2	0	–	–
3	1	2	3	2	3	0	–
1	3	1	1	1	1	3	0

Требуется определить основные социологические и политические факторы, действующие на мнение респондентов, а также исследовать взаимное положение политических лидеров в пространстве этих факторов.

Произведем непараметрическое шкалирование этих данных в евклидовой метрике с минимизацией по стрессу 1 Крассела в 2–мерном пространстве и с выдачей графика проекции на плоскость первых двух компонент.

### Результаты:

МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ. Файл: msc.std

	Собственные значения координатных осей шкалирования					
Координата:	1	2	3	4	5	6
Собств.зн	9.92	3.22	2.14	1.39		
Дисперс%	59.5	19.3	12.8	8.32		
Накоплен%	59.5	78.8	91.7	100		

Метод: Крассел1+Эвклид

Координаты объектов в результате шкалирования

Координата:	1	2	3	4	5	6
Popov	0.764	-0.651				
Gidaspov	-1.44	0.463				
Gorbi	-0.338	0.807				
Bush	0.667	0.732				
Eltsin	-0.0465	-1.08				
Lansberg	1.49	0.2				
Ligachev	-1.97	-0.382				
Sobchak	0.873	-0.0874				

Стресс 1-4 = 0.132; 1.16, 0.338, 3.54

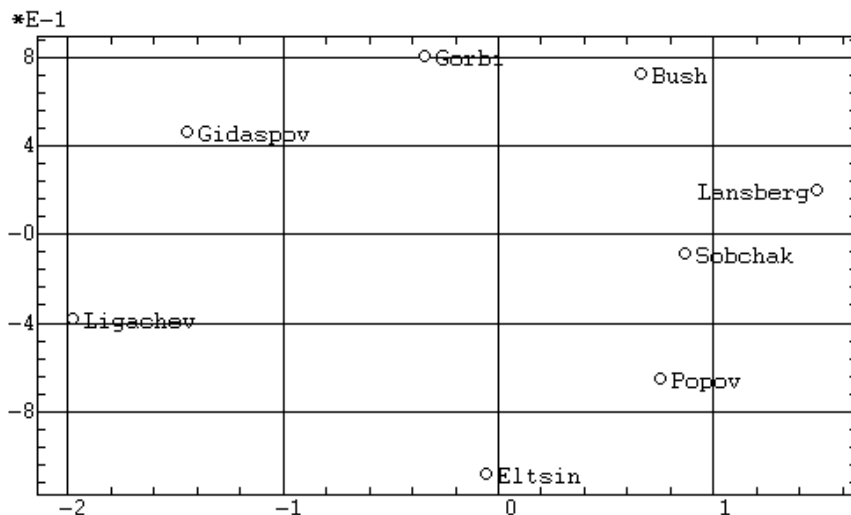


Рис. 11.32. Расположение политических лидеров 1991 г. в плоскости первых двух факторов

**Обсуждение:** Как следует из числовой выдачи, исходные данные уверенно погружаются в 4–мерное пространство, в котором первые два фактора покрывают 79% дисперсии. Проекция объектов на плоскость первых двух факторов (рис. 11.32) с учетом политической ориентации лидеров позволяет достаточно уверенно интерпретировать первую главную ось (абсцисс) как «демократичность» или «популярность», а вторую — как «признанность в качестве официального политического деятеля».

Для сравнения приведем результаты шкалирования методом Торгерсона (рис. 11.33). Заметное отличие заключается только в сближении Попова и Собчака. Отметим, что для данного примера формула оценки стресса Говера дает результаты, близкие к Торгерсону, а все три формулы стресса Краскеда дают практически идентичные результаты.

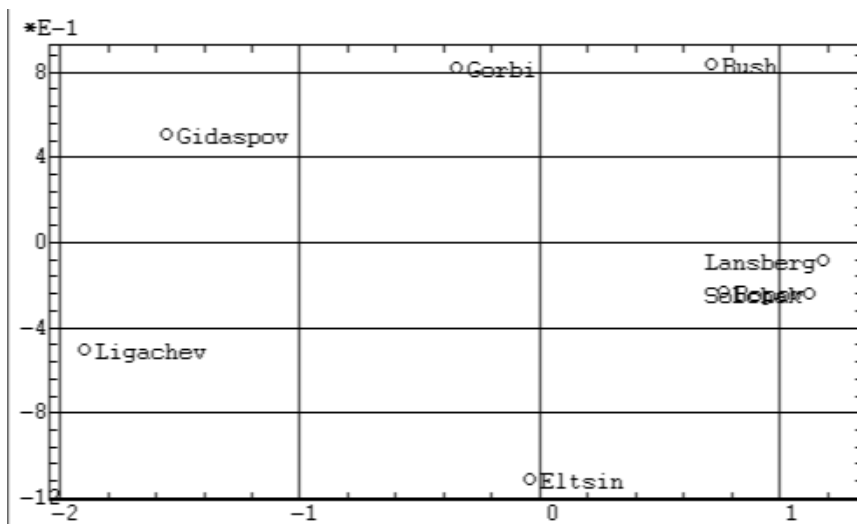


Рис. 11.33. Результаты метрического шкалирования методом Торнгенсона

## Пример 2

Проверим эффективность процедуры многомерного шкалирования на следующем наглядном примере. Возьмем карту СССР и на ней выберем координаты крупных городов. Далее измеряем взаимные расстояния между всеми парами городов и округляем их со случайными погрешностями, дабы сделать граф расстояний неплоским (табл. 11.7, файл TOWN).

Таблица 11.7. Приблизительные расстояния между городами СССР

Москва	Киев	СПБ	Архан	Саратов	Пермь	Краснод	Минск	Казань	Астрах
0	-	-	-	-	-	-	-	-	-
47	0	-	-	-	-	-	-	-	-
39	65	0	-	-	-	-	-	-	-
62	105	47	0	-	-	-	-	-	-
45	68	83	92	0	-	-	-	-	-
72	116	93	69	60	0	-	-	-	-
75	54	109	135	56	116	0	-	-	-
42	27	43	87	78	113	80	0	-	-
45	85	74	67	32	32	87	86	0	-
81	86	119	130	39	78	45	105	66	0

Обработаем полученную матрицу взаимных расстояний процедурой многомерного шкалирования и сравним полученный результат с расположением городов на карте СССР.

**Выводы:** Результаты шкалирования (рис. 11.34) вполне соответствуют реальному распределению городов на карте СССР.



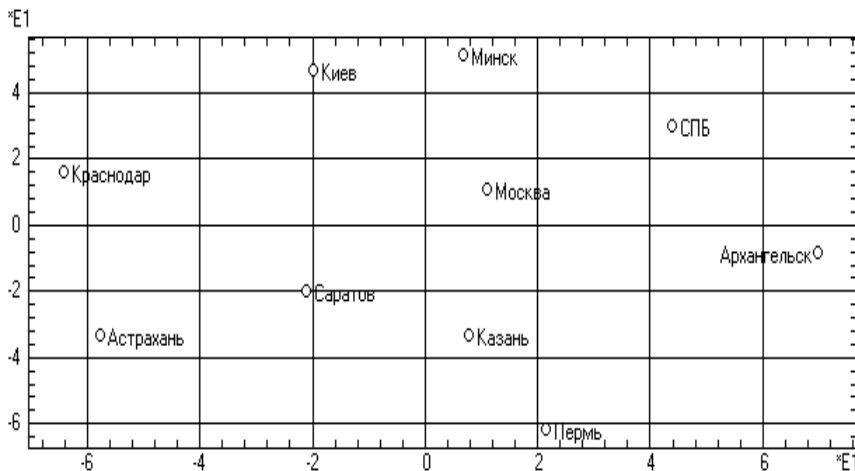


Рис.11.34. Расположение городов СССР (запад вверх)

## **ВЕРОЯТНОСТИ И ЧАСТОТЫ**

### **12.2. Вычисления вероятностей**

#### **Пример 1**

**Задача.** Вычислить 13 значений функции вероятности биномиального распределения с вероятностью успеха 0,7 и числом испытаний 10 и построить график распределения.

**Результаты:**

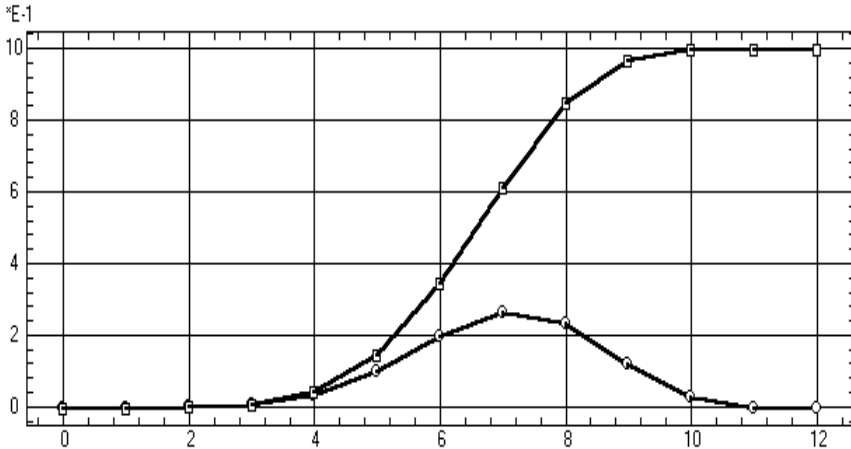


Рис. 12.5. График вероятности и плотности вероятности числа успехов  $X$  ( $X = 0, 1, 2, \dots$ ) в 10 испытаниях при вероятности успеха в одном испытании, равной 0,7 (биномиальное распределение)

ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ. Распределение биномиальное: 0.7, 10, 2  
Среднее=7, Дисперсия=2.1, Ст.отклонение=4.41

Функция распределения вероятностей

r	$P(=r)$	$P(X \leq r)$
0	5.9E-6	5.9E-6
1	0.000138	0.000144
2	0.00145	0.00159
3	0.009	0.0106
4	0.0368	0.0473
5	0.103	0.15
6	0.2	0.35
7	0.267	0.617
8	0.233	0.851
9	0.121	0.972
10	0.0282	1

## Пример 2

**Задача.** Вычислить несколько значений функции вероятности распределения Вейбулла с параметрами: масштаба, равным 5, и формы, равным 3, и построить график распределения.

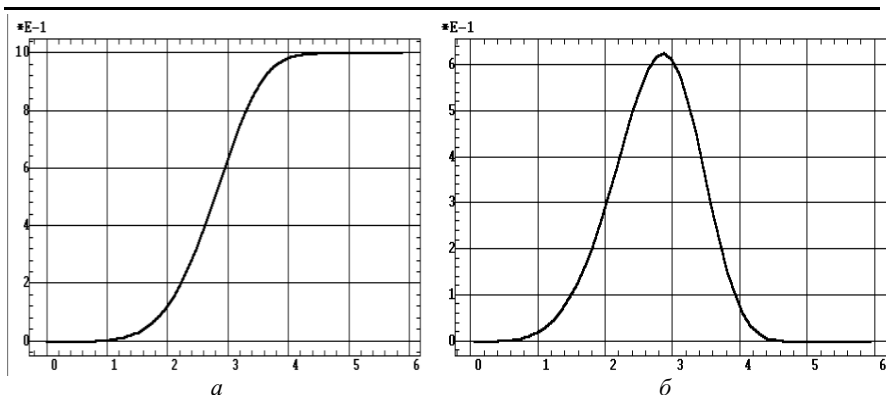


Рис. 12.6. Графики распределения Вейбулла с параметрами: формы, равным 3, и масштаба, равным 5:

$a$  — функции вероятности;  $b$  — функции плотности вероятности

### Результаты:

ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ. Распределение Вейбулла: 2, 3

Среднее=1.79, Дисперсия=2.81, Ст.отклонение=7.92

x1	F(x1)	x2	F(x2)	F(x1)-F(x2)
1	2	0.118	0.632	0.515
2	3	0.632	0.966	0.334
3	4	0.966	1	0.0339
4	5	1	1	0.000335
5	6	1	1	1.64E-7
6	7	1	1	1.82E-12
7	8	1	1	0

## 12.3. Согласие распределений

### Пример 1

**З а д а ч а.** Выполнить проверку соответствия теоретического распределения и эмпирического распределения выборки объемом 30 чисел, полученной генерацией по распределению Вейбулла с параметрами: масштаба, равным 2, и параметром формы, равным 3, и выдать совместный график теоретического и эмпирического распределений.

### Результаты:

СОГЛАСИЕ РАСПРЕДЕЛЕНИЙ. Распределение Вейбулла: 1.98, 3.67

Колмогоров=0.0666, Значимость=1, степ.своб = 30

Гипотеза 0: <Распределение не отличается от теоретического>

Омега-квадрат=0.016, Значимость=1, степ.своб = 30

Гипотеза 0: <Распределение не отличается от теоретического>

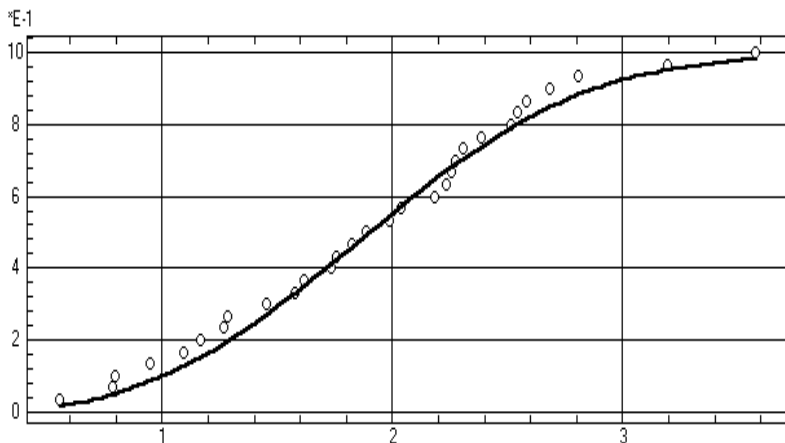


Рис. 12.8. Графики эмпирического и теоретического (Вейбулла) распределений

## 12.4. Согласие частот событий (долей)

---

### Пример 1

**Задача.** В одной выборке объема 1000 оказалось 20 бракованных изделий, в другой выборке объема 900 выявлено 30 бракованных изделий. Необходимо проверить, совпадают ли частоты появления бракованных изделий. Большое число наблюдений позволяет использовать  $Z$ -статистику.

#### Результаты:

СОГЛАСИЕ ЧАСТОТ.

$Z = -1.956$ , значимость =  $5.04E-2$

Гипотеза 0: <Частоты событий совпадают>

Разность частот = 0.0133, доверит. интервал = 0.0121

**Выводы:** Результаты анализа не позволяют принять ни одну из гипотез о соотношении частот появления брака (уровень значимости близок к критическому значению 0,05), поэтому требуется накопление дополнительных данных. Это подтверждает и противоположный вывод по сравнению разности частот с доверительным интервалом, который, несмотря на близость к разности частот, не включает нулевое значение.

## Пример 2

**Задача.** Продолжим анализ соотношений встречаемости тромбоза при приеме аспирина и контроле из примера 1 разд. 7.6: при приеме аспирина наблюдалось 6 случаев тромбоза у 19 больных, а в контрольной группе — 18 случаев у 25 больных. Большое число наблюдений позволяет использовать  $Z$ -статистику.

### Результаты:

СОГЛАСИЕ ЧАСТОТ.

$Z=2.8$ , Значимость=0.00508

Гипотеза 1: <Частоты событий не совпадают>

Разность частот=0.377, доверит.интервал=0.248

**Выводы:** Результаты анализа позволяют с достаточной уверенностью ( $P=0,0029$  значительно ниже критического значения 0,05, а доверительный интервал относительно разности частот не включает нулевое значение) принять гипотезу о различии этих двух частот, что подтверждает результаты примера 1 из разд. 7.6.

## 12.5. Последовательный анализ

### Пример

**Задача.** Было установлено, что затраты на предпосевную обработку семян окупятся, если прирост урожайности составит не менее  $d=3$  ц/га при средней урожайности необработанных семян  $x_0=20,8$  ц/га. Многолетние наблюдения показали, что в норме  $S_x=1,81$  ц/га. В табл. 12.1 приведены урожайности обработанных семян за 10 последовательных лет (файл SEQ).

Таблица 12.1. Урожайность зерновых при предпосевной обработке семян

1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
26.7	21	24.1	27.1	25.1	23	26.2	19.4	21.8	23.4

Необходимо проверить, достаточно ли произведенных оценок урожайности (испытаний) для подтверждения гипотезы об окупаемости затрат.

**Выводы:** Как показывает график результатов последовательного анализа, оценки урожайности постепенно и уверенно выходят за

допусковые границы, поэтому можно принять альтернативную гипотезу и закончить испытания уже на 5–6 году.

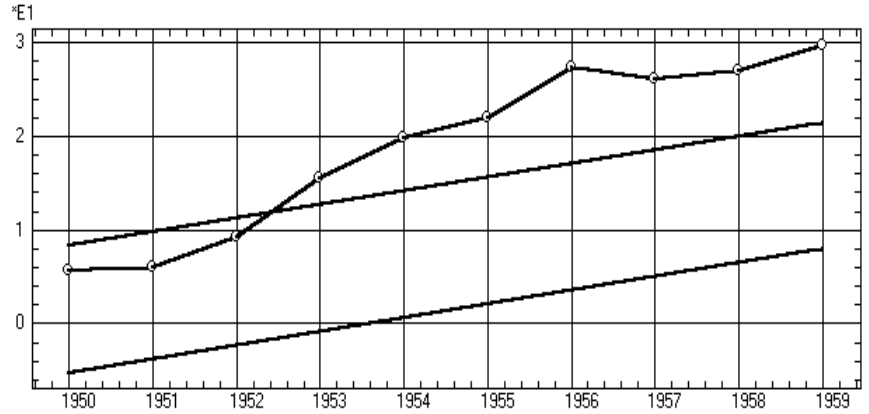


Рис. 12.11. График последовательного анализа эффективности предпосевной обработки семян с зоной принятия нулевой гипотезы

## 12.6. Анализ выживаемости

### Пример

**З а д а ч а.** При остром лимфобластном лейкозе мутация предшественника лимфоцитов приводит к появлению клона лейкозных клеток, способных неограниченно делиться. Они подавляют нормальное кроветворение, вызывая иммунодефицит, анемию и тромбоцитопению, приводящие к смерти. При лечении облучением и химиотерапией для компенсации их побочных действий используют пересадку костного мозга. Для избежания отторжения костной ткани лучше использовать косный мозг близких родственником (аллотрансплантация). Однако не у всех таковые имеются или согласны на донорство, поэтому используется и пересадка собственного костного мозга (аутотрансплантация).

В табл. 12.2 приведены данные о числе умерших и выбывших из наблюдения пациентов (столбцы «выбыли») при этих двух способах пересадки костного мозга (файл LIFE2). В колонках «время» указаны месяцы исходов.

Таблица 12.2. Данные выживаемости пациентов с лейкозом при аутотрансплантации (1) и аллотрансплантации (2) костного мозга

Время1	Выбыли1	Время2	Выбыли2
1	3	1	1
2	2	2	1
3	1	3	1
4	1	4	1

5	1	6	1
6	1	7	1
7	1	12	1
10	1	20	-1
12	2	21	-1
14	1	24	1
17	1	30	-1
20	-1	60	-1
27	2	85	-2
28	1	86	-1
30	2	87	-1
36	1	90	-1
38	-1	100	-1
40	-1	119	-1
45	-1	132	-1
50	3		
63	-1		
132	-2		

Возникает задача сравнения этих двух видов трансплантации по степени выживаемости пациентов.

**Обсуждение:**

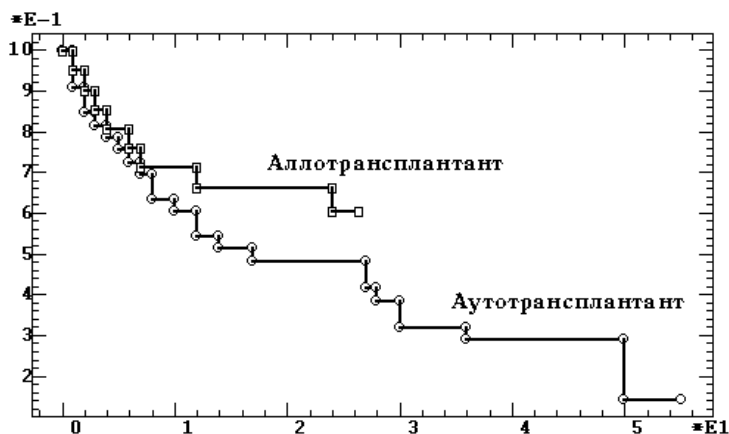


Рис. 12.13. Сравнение двух кривых выживаемости

СРАВНЕНИЕ ВЫЖИВАЕМОСТЕЙ. Файл: life2.std

Время	Пережили	Выжили	Ст.ошибка	Время	Пережили	Выжили	Ст.ошибка
1	0.909	0.909	0,05	1	0.952	0.952	0.0465
2	0.933	0.848	0.0624	2	0.95	0.905	0.0641
3	0.964	0.818	0.0671	3	0.947	0.857	0.0764
4	0.963	0.788	0.0712	4	0.944	0.81	0.0857
5	0.962	0.758	0.0746	6	0.941	0.762	0.0929
6	0.96	0.727	0.0775	7	0.938	0.714	0.0986
7	0.958	0.697	0.08	12	0.933	0.667	0.103
8	0.913	0.636	0.0837	24	0.909	0.606	0.11
10	0.952	0.606	0.0851				
12	0.9	0.545	0.0867				
14	0.944	0.515	0.087				
17	0.941	0.485	0.087				

27	0.867	0.42	0.0866
28	0.923	0.388	0.0857
30	0.833	0.323	0.0827
36	0.9	0.291	0.0805
50	0.5	0.145	0.0717

Z=2.16, значимость=0.0153  
 Гипотеза 1: <Есть различия между кривыми выживаемости>

**В ы в о д ы:** Как показывают результаты статистического сравнения, две кривые выживаемости различаются на высоком уровне достоверности ( $P=0,0153$ ). Поэтому можно обратиться к визуальному исследованию совместного графика кривых выживаемости (рис. 12.13). На графике видно, что аллотрансплантация костного мозга имеет несомненное преимущество, способствуя значительно большей выживаемости пациентов. Медиана аутооттрансплантации находится в районе 17-го месяца, а к 50-му месяцу выживает только 14,5% пациентов. При аллотрансплантации медиана экспериментально не достигается (ее можно визуально прогнозировать в районе 50 месяцев), и как видно из табл. 12.2 — за счет пациентов, выбывших из-под наблюдения, по всей видимости вследствие стабилизации их состояния.

# МЕТОДЫ КОНТРОЛЯ КАЧЕСТВА

## 13.1. Гистограмма качества

### Пример

**Д а н н ы е:** При выходном контроле были измерены диаметры 125 последовательно изготовленных деталей — цапф. Анализу были подвергнуты две последние значащие цифры проведенных измерений (переменная (x-R) файла QCONTR).

47	32	44	35	20	19	37	31	25	34	19	11	16	11	44	29	29	42	59	38	28	12
45	36	25	40	35	11	38	33	15	30	12	33	26	35	44	32	11	38	27	37	26	20
35	23	45	26	37	32	28	44	40	31	18	31	25	24	32	22	22	37	19	47	14	37
32	12	38	30	25	40	24	50	19	7	31	23	18	32	38	0	41	40	37	35	12	29
48	20	31	20	35	24	47	12	27	38	40	31	52	42	52	24	25	20	31	15	3	28
29	47	41	32	22	28	27	22	32	54	42	34	15	29	21							

### Результаты:

X	X-станд	Частота	%	Накопл.	%
0	-2.636	3	2.4	3	2.4
7.375	-1.985	10	8	13	10.4
14.75	-1.334	20	16	33	26.4
22.13	-0.6831	26	20.8	59	47.2
29.5	-3.213E-2	28	22.4	87	69.6
36.88	0.6188	26	20.8	113	90.4
44.25	1.27	8	6.4	121	96.8
51.63	1.921	4	3.2	125	100
59	2.572				



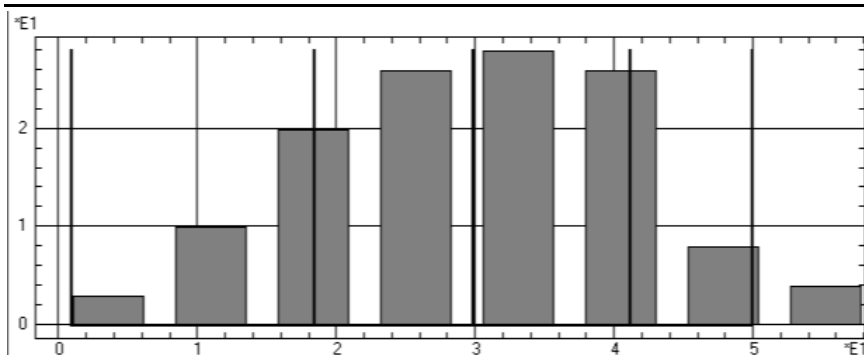


Рис. 13.1. Гистограмма для измерений диаметров цапф, вертикальными линиями указаны: среднее значение, стандартное отклонение и границы допуска

**В ы в о д ы:** Как видно из рис. 13.1 распределение диаметров цапф достаточно симметрично с небольшим смещением в область больших диаметров, но значительное число изделий выходит за границы допуска.

## 13.2. Диаграмма Парето

### Пример

**Д а н н ы е:** При выходном контроле было подсчитано число дефектов следующих последовательных категорий (переменная DIAGR файла QCONTR): деформации=104, царапины=42, раковины=20, трещины=10,

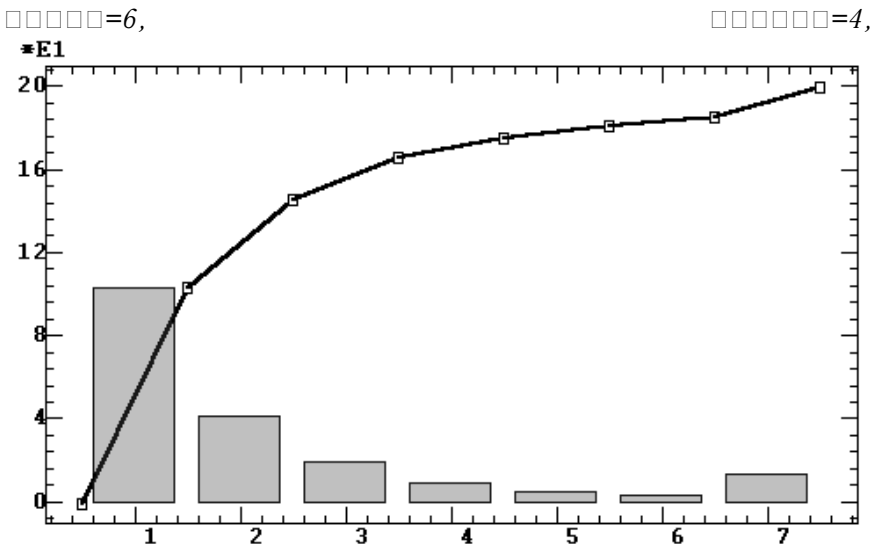


Рис. 13.2. Диаграмма Парето для дефектов в готовой продукции

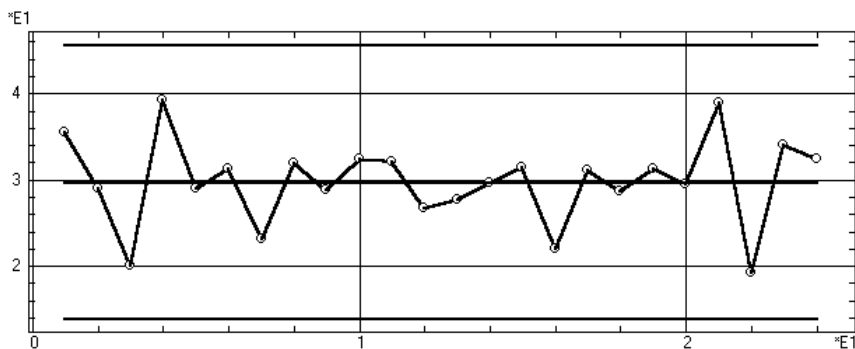
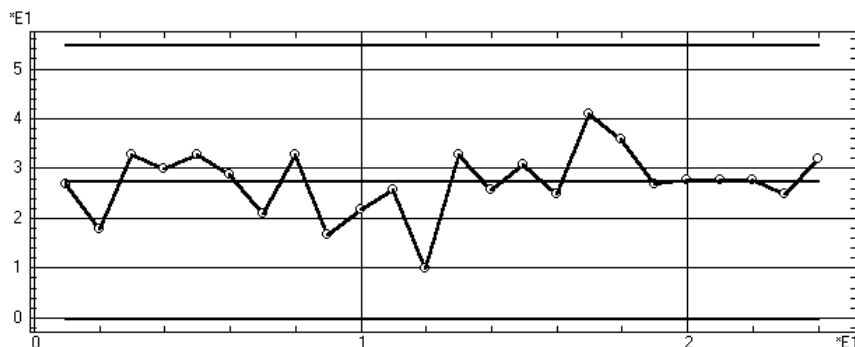
**В ы ы в о д ы:** Как видно из рис. 13.2 подавляющее число дефектов приходится на деформации (именно на их уменьшение необходимо в первую очередь обратить внимание технологов), количество же трещин, пятен и разрывов сравнительно невелико и в сумме сопоставимо с прочими дефектами.

### 13.3. Контрольные карты

---

#### Примеры

**Д а н н ы е 1:** При выходном контроле были измерены диаметры 125 последовательно изготовленных деталей — цапф (см. пример к предыдущему разделу). Анализу были подвергнуты две последние значащие цифры проведенных измерений:

Рис. 13.4. Контрольная  $\bar{x}$ -карта для измерений диаметра цапфРис. 13.5. Контрольная  $R$ -карта для измерений диаметра цапф

**В ы в о д ы:** Полученные по этим данным  $(\bar{x}-R)$ -контрольные карты (рис. 13.4, 13.5) не выходят за границы допуска, т. е. не обнаруживают тенденций, требующих управления.

**Д а н н ы е 2:** При выходном контроле было подсчитано дефектных изделий в последовательных партиях по 100 изделий в каждой (переменная PN файла QCONTR):

4 2 0 5 3 2 4 3 2 6 1 4 1 0 2 3 1 6 1 3 3 2 0 7 3

**В ы в о д ы:** Согласно  $pn$ -контрольной карте (рис. 13.6) процесс находится в допустимых границах.

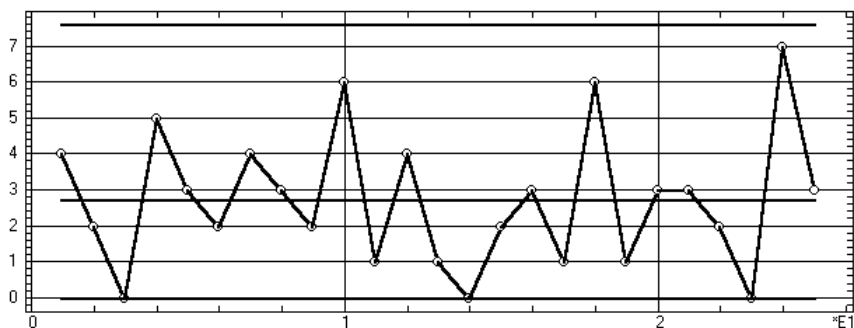


Рис. 13.6. Контрольная  $p$ -карта для числа дефектных изделий

**Д а н н ы е 3:** При выходном контроле было подсчитано количество дефектных изделий  $n$  в последовательных партиях разного размера  $L$  (переменные P1 и P2 файла QCONTR):

No.	L	n	No.	L	n	No.	L	n	No.	L	n	No.	L	n
1	1500	46	2	1850	44	3	1902	84	4	985	14	5	1600	55
6	1550	116	7	1600	134	8	1650	114	9	1800	119	10	1100	61
11	1450	90	12	1900	91	13	2000	38	14	1900	56	15	1650	45
16	1200	23	17	1700	56	18	1850	41	19	2000	22	20	1800	34
21	1780	32	22	1050	17	23	1750	65	24	1850	61	25	1900	72

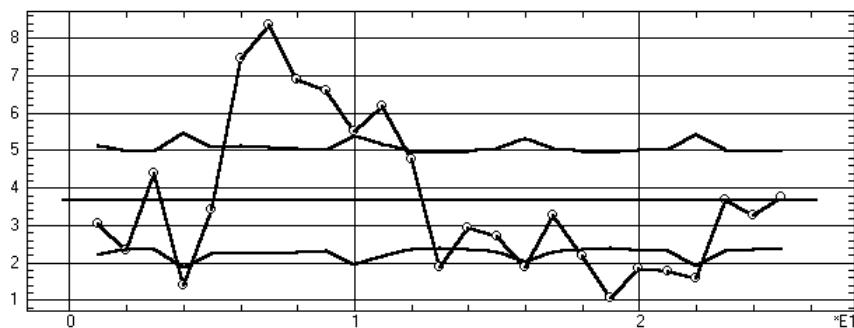


Рис. 13.7. Контрольная  $p$ -карта для числа дефектных изделий

**В ы в о д ы:** Анализ полученной  $p$ -контрольной карты (рис. 13.7) выявляет резкий скачок числа дефектных изделий в партиях с шестой по двенадцатую. По-видимому, начиная с 12-й партии было применено управление, которое позволило улучшить качество продукции.

**Д а н н ы е 4:** При выходном контроле было подсчитано число утолщений  $n$  в волокнах различного размера  $L$  [тыс.м] (переменные U1 и U2 файла QCONTR):

No.	L	n	No.	L	n	No.	L	n	No.	L	n	No.	L	n
1	4	3	2	5	6	3	4.5	17	4	4	2	5	6.2	3
6	4	5	7	3	2	8	6	1	9	4	1	10	4	2

11	4	1	12	4	1	13	4	4	14	5.5	2	15	6.5	5
16	6.5	2	17	6	1	18	4	3	19	4	1	20	4	1

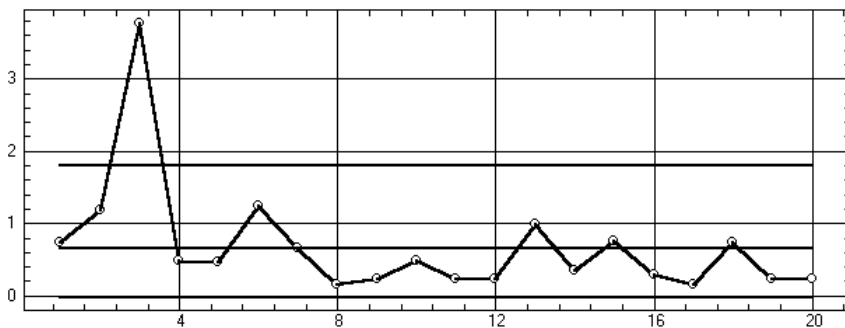


Рис. 13.8. Контрольная  $u$ -карта для числа утолщений в волокне

**В ы в о д ы:** На полученной  $u$ -контрольной карте (рис. 13.8) выделяется плохое качество волокон в третьей пробе.