Глава 3
Информационное моделирование

§ 16
Компьютерное информационное моделирование

Известно, что модель — это некоторое упрощенное подобие реального объекта. Более полное определение звучит так:

Модель — это объект-заменитель, который в определенных условиях может заменять объект-оригинал. Модель воспроизводит интересующие нас свойства и характеристики оригинала.

Модели бывают материальными и информационными. Примерами материальных моделей являются глобус — модель Земли; манекен — модель человеческого тела; модели самолетов, кораблей, ракет, автомобилей; макет застройки жилого района в городе и многое другое.

Предметом изучения информатики являются информационные модели.

В **информационной модели** отражаются знания человека об объекте моделирования. Информационная модель — это описание в той или иной форме объекта моделирования.

Объектом информационного моделирования может быть всё, что угодно: отдельные предметы (дерево, стол); физические, химические, биологические процессы (течение воды в трубе, получение серной кислоты, фотосинтез в листьях растений); метеорологические явления (гроза, смерч); экономические и социальные процессы (динамика цен акций на бирже, миграция населения).

Можно сказать, что информационным моделированием занимается любая наука, поскольку задача науки состоит в получении знаний, а наши знания о действительности всегда носят приближенный, т. е. *модельный*, характер. С развитием науки эти знания уточняются, углубляются, но всё равно остаются приближенными. Старые модели заменяются на новые, более точные, и этот процесс бесконечен.

Физика создает модели физических объектов, химия — химических, экономика и социология — социально-экономических и т. д.

Информатика занимается общими методами и средствами создания и использования информационных моделей.

Компьютерная информационная модель. Основным инструментом современной информатики является компьютер. Поэтому информационное моделирование в информатике — это компьютерное моделирование, применимое к объектам различных предметных областей. Компьютер позволил ученым работать с такими информационными моделями, исследование которых было невозможно или затруднено в докомпьютерные времена. Например, метеорологи могли и 100 лет назад написать уравнения для расчета прогноза погоды на завтра. Но на решение их «ручным способом» потребовалось бы много лет. И лишь с помощью компьютера появилась возможность рассчитать прогноз погоды прежде, чем наступит завтрашний день.

Чаще всего информационное моделирование используется для прогнозирования поведения объекта моделирования, для принятия управляющих решений. Характерной особенностью компьютерных информационных моделей является возможность их использования в режиме реального времени, т. е. с соблюдением временны́х ограничений на получение результата. В самом деле, какой смысл имеет получение через неделю прогноза на завтра или расчет управляющего решения через час, если его принятие требуется через пять минут? Высокое быстродействие современных компьютеров снимает эти проблемы.

Этапы моделирования (рис. 3.1). Построение информационной модели начинается с системного анализа объекта моделирования. Представим себе быстро растущую фирму, руководство которой столкнулось с проблемой снижения эффективности работы фирмы по мере ее роста (что является обычной ситуацией) и решило упорядочить управленческую деятельность. Первое, что

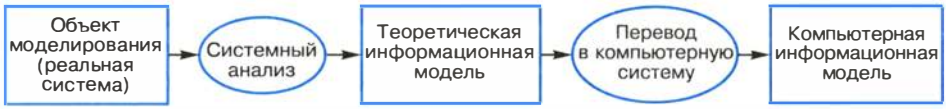

Рис. 3.1. Этапы разработки компьютерной информационной модели

будет сделано на этом пути, — *системный анализ* деятельности фирмы, т. е. анализ объекта моделирования как системы в соответствии с системным подходом (см. § 1). Системный аналитик, приглашенный в фирму, должен изучить ее деятельность, выделить участников процесса управления и их деловые взаимоотношения.

Далее полученное теоретическое описание моделируемой системы преобразуется в компьютерную модель. Для этого либо используется готовое программное обеспечение, либо привлекаются программисты для его разработки. В конечном итоге получается компьютерная информационная модель, которая будет использоваться по своему назначению.

Для нашего примера с фирмой компьютерная информационная модель поможет найти оптимальный вариант управления, при котором будет достигнута наивысшая эффективность работы фирмы согласно заложенному в модель критерию (например, это может быть максимум прибыли на единицу вложенных средств).

Информационная модель базируется на данных, т. е. на информации об объекте моделирования. Любой реальный объект обладает бесконечным множеством различных свойств. Для создания его информационной модели требуется выделить лишь те свойства, которые необходимы с точки зрения цели моделирования; четко сформулировать эту цель необходимо до начала моделирования. Например, если вы хотите создать модель учебного процесса в вашем классе, то вам потребуются данные об изучаемых предметах, расписании занятий, оценках учеников, преподавателях. А если вы захотите смоделировать процесс летнего отдыха (например, коллективной поездки на юг), то вам потребуются совсем другие данные: сроки поездки, маршрут поезда, стоимость билетов, стоимость расходов на питание и пр. Возможно, что единственными общими данными для этих двух моделей будет список учеников класса.

Система основных понятий

<table>
<tr><td colspan="4">Компьютерное информационное моделирование</td></tr>
<tr><td colspan="4">Модель — это объект-заменитель реального объекта</td></tr>
<tr><td colspan="4">Виды моделей:</td></tr>
<tr><td rowspan="4">Материальные (натурные) модели</td><td colspan="3">Информационные модели</td></tr>
<tr><td colspan="3">Компьютерная информационная модель — модель, реализованная на компьютере</td></tr>
<tr><td colspan="3">Этапы построения компьютерной информационной модели:</td></tr>
<tr><td>Определение цели моделирования</td><td>Системный анализ объекта моделирования: результат — теоретическая информационная модель</td><td>Реализация модели на компьютере: используется специальное программное обеспечение или языки высокого уровня</td></tr>
</table>

Вопросы и задания

1. Что такое модель? Приведите примеры материальных моделей, не упомянутых в параграфе.
2. Что такое информационная модель?
3. Можно ли карту города назвать информационной моделью? Обоснуйте ответ.
4. Почему многие научные знания можно отнести к информационным моделям?
5. Какова роль информатики в информационном моделировании?
6. В чем преимущество компьютерных информационных моделей перед теоретическими?
7. Какие данные вы бы включили в информационные модели следующих объектов и процессов:
   - обед в школьной столовой;
   - ремонт квартиры;
   - пассажир поезда;
   - дом, в котором вы живете?

§ 17
Моделирование зависимостей между величинами

Величины и зависимости между ними

Содержание данного раздела учебника связано с компьютерным математическим моделированием. Применение математического моделирования постоянно требует учета зависимостей одних величин от других. Приведем примеры таких зависимостей:

1) время падения тела на землю зависит от его первоначальной высоты;
2) давление газа в баллоне зависит от его температуры;
3) уровень заболеваемости жителей города бронхиальной астмой зависит от концентрации вредных примесей в городском воздухе.

Реализация математической модели на компьютере (*компьютерная математическая модель*) требует владения приемами представления зависимостей между величинами.

Рассмотрим различные методы представления зависимостей.

Всякое исследование нужно начинать с выделения количественных характеристик исследуемого объекта. Такие характеристики называются **величинами**.

С понятием величины вы уже встречались в курсе информатики 7–9 классов. Напомним, что со всякой величиной связаны три основных свойства: *имя, значение, тип*.

Имя величины может быть смысловым и символическим. Примером смыслового имени является «давление газа», а символическое имя для этой же величины — P. В базах данных величинами являются поля записей. Для них, как правило, используются смысловые имена, например: ФАМИЛИЯ, ВЕС, ОЦЕНКА и т. п. В физике и других науках, использующих математический аппарат, применяются символические имена для обозначения величин. Чтобы не терялся смысл, для определенных величин используются стандартные имена. Например, время обозначают буквой t, скорость — V, силу — F и пр.

Если значение величины не изменяется, то она называется постоянной величиной или **константой**. Пример константы — число Пифагора $\pi = 3{,}14159\ldots$. Величина, значение которой может меняться, называется **переменной**. Например, в описании процес-

са падения тела переменными величинами являются высота H и время падения t.

Третьим свойством величины является ее **тип**. С понятием типа величины вы также встречались, знакомясь с программированием и базами данных. Тип определяет множество значений, которые может принимать величина. Основные типы величин: числовой, символьный, логический. Поскольку в данном разделе мы будем говорить лишь о количественных характеристиках, и рассматриваться будут только величины числового типа.

А теперь вернемся к примерам 1–3 (см. начало параграфа) и обозначим (поименуем) все переменные величины, зависимости между которыми нас будут интересовать. Кроме имен укажем размерности величин. Размерности определяют единицы, в которых представляются значения величин.

1) t (с) — время падения; H (м) — высота падения. Зависимость будем представлять, пренебрегая учетом сопротивления воздуха; ускорение свободного падения g (м/с^2) будем считать константой.
2) P (н/м^2) — давление газа (в единицах СИ давление измеряется в ньютонах на квадратный метр); t (°С) — температура газа. Давление при нуле градусов P_0 будем считать константой для данного газа.
3) Загрязненность воздуха будем характеризовать концентрацией примесей (каких именно, будет сказано позже) — C (мг/м^3). Единица измерения — масса примесей, содержащихся в 1 кубическом метре воздуха, выраженная в миллиграммах. Уровень заболеваемости будем характеризовать числом хронических больных астмой, приходящихся на 1000 жителей данного города — P (бол./тыс.).

Отметим важное качественное различие между зависимостями, описанными в примерах 1 и 2, с одной стороны, и в примере 3, с другой. В первом случае зависимость между величинами является полностью определенной: значение H однозначно определяет значение t (пример 1), значение t однозначно определяет значение P (пример 2). Но в третьем примере зависимость между значением загрязненности воздуха и уровнем заболеваемости носит существенно более сложный характер; при одном и том же уровне загрязненности в разные месяцы в одном и том же городе (или в разных городах в один и тот же месяц) уровень заболеваемости может быть разным, поскольку на него влияют и многие

другие факторы. Отложим более детальное обсуждение этого примера до следующего параграфа, а пока лишь отметим, что на математическом языке зависимости в примерах 1 и 2 являются **функциональными**, а в примере 3 — нет.

Математические модели

Если зависимость между величинами удается представить в математической форме, то мы имеем математическую модель.

Математическая модель — это совокупность количественных характеристик некоторого объекта (процесса) и связей между ними, представленных на языке математики.

Хорошо известны математические модели для первых двух примеров. Они отражают физические законы и представляются в виде формул:

$$t = \sqrt{\frac{2H}{g}}; \quad P = P_0\left(1 + \frac{t}{273}\right).$$

Это примеры зависимостей, представленных в функциональной форме. Первую зависимость называют корневой (время пропорционально квадратному корню высоты), вторую — линейной.

В более сложных задачах математические модели представляются в виде уравнений или систем уравнений. В конце данной главы будет рассмотрен пример математической модели, которая выражается системой неравенств.

В еще более сложных задачах (пример 3 — одна из них) зависимости тоже можно представить в математической форме, но не функциональной, а иной.

Табличные и графические модели

Рассмотрим примеры двух других, не формульных, способов представления зависимостей между величинами: **табличного** и **графического**. Представьте себе, что мы решили проверить закон свободного падения тела экспериментальным путем. Эксперимент организуем следующим образом: будем бросать стальной шарик с 6-метровой высоты, 9-метровой и т. д. (через 3 метра), замеряя высоту начального положения шарика и время падения. По результатам эксперимента составим таблицу и нарисуем график (рис. 3.2).

H, м	t, с
6	1,1
9	1,4
12	1,6
15	1,7
18	1,9
21	2,1
24	2,2
27	2,3
30	2,5

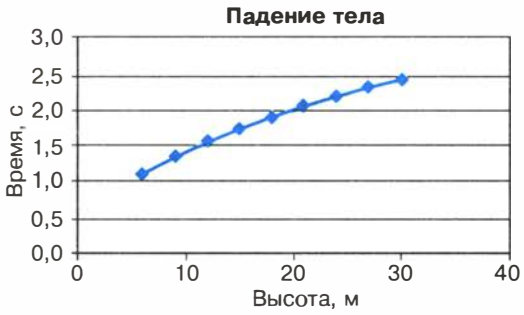

Рис. 3.2. Табличное и графическое представление зависимости времени падения тела от высоты

Если каждую пару значений H и t из данной таблицы подставить в приведенную выше формулу зависимости времени от высоты, то формула превратится в равенство (с точностью до погрешности измерений). Значит, модель работает хорошо. (Однако если сбрасывать не стальной шарик, а большой легкий мяч, то равенство не будет достигаться, а если надувной шарик, то значения левой и правой частей формулы будут различаться очень сильно. Как вы думаете почему?)

В этом примере мы рассмотрели три способа моделирования зависимости величин: функциональный (формула), табличный и графический. Однако *математической моделью* процесса падения тела на землю можно назвать только формулу. Формула более универсальна, она позволяет определить время падения тела с любой высоты, а не только для того экспериментального набора значений H, который отображен на рис. 3.2. Имея формулу, можно легко создать таблицу и построить график, а наоборот — весьма проблематично.

Точно так же тремя способами можно отобразить зависимость давления от температуры. Оба примера связаны с известными физическими законами — законами природы. Знания физических законов позволяют производить точные расчеты, они лежат в основе современной техники.

Информационные модели, которые описывают развитие систем во времени, имеют специальное название: **динамические модели**. В примере 1 приведена именно такая модель. В физике динамические информационные модели описывают движение тел,

в биологии — развитие организмов или популяций животных, в химии — протекание химических реакций и т. д.

Система основных понятий

<table>
<tr><td colspan="4">Моделирование зависимостей между величинами</td></tr>
<tr><td colspan="4">Величина — количественная характеристика исследуемого объекта</td></tr>
<tr><td colspan="4">Характеристики величины</td></tr>
<tr><td rowspan="2">Имя:
отражает смысл величины</td><td rowspan="2">Тип:
определяет возможные значения величины</td><td colspan="2">Значение</td></tr>
<tr><td>константа</td><td>переменная</td></tr>
<tr><td colspan="4">Виды зависимостей:</td></tr>
<tr><td colspan="2">Функциональные</td><td colspan="2">Иные</td></tr>
<tr><td colspan="4">Способы отображения зависимостей</td></tr>
<tr><td>Математическая модель</td><td colspan="2">Табличная модель</td><td>Графическая модель</td></tr>
<tr><td colspan="4">Описание развития системы во времени — динамическая модель</td></tr>
</table>

Вопросы и задания

1. а) Какие вам известны формы представления зависимостей между величинами?
   б) Что такое математическая модель?
   в) Может ли математическая модель включать в себя только константы?
2. Приведите пример известной вам функциональной зависимости (формулы) между характеристиками какого-то объекта или процесса.
3. Обоснуйте преимущества и недостатки каждой из трех форм представления зависимостей.

§ 18
Модели статистического прогнозирования

О статистике и статистических данных

Рассмотрим способ нахождения зависимости частоты заболеваемости жителей города бронхиальной астмой от качества воздуха (третий пример из сформулированных в начале предыдущего параграфа). Любому человеку понятно, что такая зависимость существует. Очевидно, что чем хуже воздух, тем больше больных астмой. Но это качественное заключение. Его недостаточно для того, чтобы управлять уровнем загрязненности воздуха. Для управления требуются более конкретные знания. Нужно установить, какие именно примеси сильнее всего влияют на здоровье людей, как связана концентрация этих примесей в воздухе с числом заболеваний. Такую зависимость можно установить только экспериментальным путем: посредством сбора многочисленных данных, их анализа и обобщения.

При решении таких проблем на помощь приходит статистика.

Статистика — наука о сборе, измерении и анализе массовых количественных данных.

Существуют медицинская статистика, экономическая статистика, социальная статистика и другие. Математический аппарат статистики разрабатывает наука под названием математическая статистика.

Рассмотрим пример из области медицинской статистики.

Известно, что наиболее сильное влияние на бронхиально-легочные заболевания оказывает угарный газ — монооксид углерода. Поставив цель определить эту зависимость, специалисты по медицинской статистике проводят сбор данных. Они собирают сведения из разных городов о средней концентрации угарного газа в атмосфере и о заболеваемости астмой (число хронических больных на 1000 жителей). Полученные данные можно свести в таблицу, а также представить в виде точечной диаграммы (рис. 3.3[1]).

[1] Приведенные в примере данные не являются официальной статистикой, однако правдоподобны.

C, мг/м^3	P, бол./тыс.
2	19
2,5	20
2,9	32
3,2	34
3,6	51
3,9	55
4,2	90
4,6	108
5	171

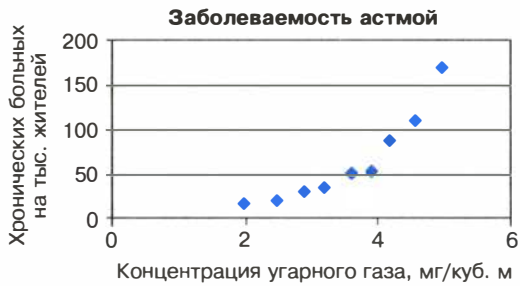

Рис. 3.3. Табличное и графическое представление статистических данных

Статистические данные всегда являются приближенными, усредненными. Поэтому они носят оценочный характер, но верно отражают характер зависимости величин. И еще одно важное замечание: для достоверности результатов, полученных путем анализа статистических данных, этих данных должно быть много.

Из полученных данных можно сделать вывод, что при концентрации угарного газа до 3 мг/м^3 его влияние на заболеваемость астмой несильное. С дальнейшим ростом концентрации наступает резкий рост заболеваемости.

А как построить математическую модель данного явления? Очевидно, нужно получить формулу, отражающую зависимость количества хронических больных P от концентрации угарного газа C. На языке математики это называется функцией зависимости P от C: $P(C)$. Вид такой функции неизвестен, ее следует искать методом подбора по экспериментальным данным.

Понятно, что график искомой функции должен проходить близко к точкам диаграммы экспериментальных данных. Строить функцию так, чтобы ее график точно проходил через все данные точки (рис. 3.4, *а*), не имеет смысла. Во-первых, математический вид такой функции может оказаться слишком сложным. Во-вторых, уже говорилось о том, что экспериментальные значения являются приближенными.

Отсюда следуют основные требования к искомой функции:

- она должна быть достаточно простой для использования ее в дальнейших вычислениях;

- график этой функции должен проходить вблизи экспериментальных точек так, чтобы отклонения этих точек от графика были минимальны и равномерны (рис. 3.4, *б*).

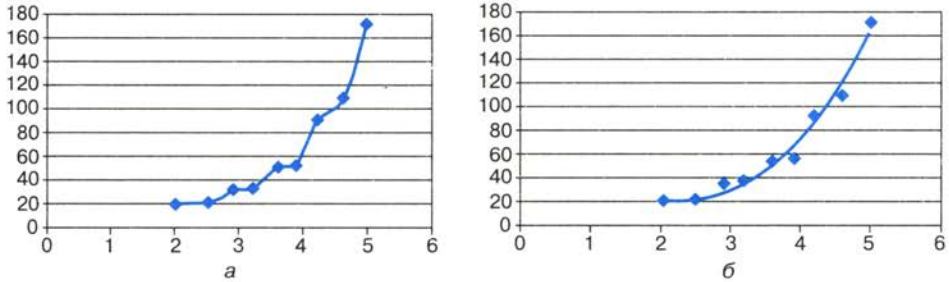

Рис. 3.4. Два варианта построения графической зависимости по экспериментальным данным

Полученную функцию, график которой приведен на рис. 3.4, *б*, в статистике принято называть **регрессионной моделью.**

Метод наименьших квадратов

Получение регрессионной модели происходит в два этапа:

1) подбор вида функции;

2) вычисление параметров функции.

Первая задача не имеет строгого решения. Здесь может помочь опыт и интуиция исследователя, а возможен и «слепой» перебор из конечного числа функций и выбор лучшей из них.

Чаще всего выбор производится среди следующих функций:

$y = ax + b$ — линейная функция;

$y = ax^2 + bx + c$ — квадратичная функция;

$y = a \ln(x) + b$ — логарифмическая функция;

$y = ae^{bx}$ — экспоненциальная функция;

$y = ax^b$ — степенная функция.

Квадратичная функция называется в математике *полиномом второй степени*. Иногда используются полиномы и более высоких степеней, например полином третьей степени имеет вид: $y = ax^3 + bx^2 + cx + d$.

Во всех этих формулах x — аргумент, y — значение функции, a, b, c, d — параметры функции, $\ln(x)$ — натуральный логарифм, e — константа, основание натурального логарифма.

Если вы выбрали (сознательно или наугад) одну из предлагаемых функций, то далее нужно подобрать параметры (a, b, c и пр.) так, чтобы функция располагалась как можно ближе к экспериментальным точкам. Что значит «располагалась как можно ближе»? Ответить на этот вопрос значит предложить метод вычисления параметров. Такой метод был предложен в XVIII веке немецким математиком К. Гауссом и называется **методом наименьших квадратов** (МНК). Суть его заключается в следующем: искомая функция должна быть построена так, чтобы сумма квадратов отклонений y-координат всех экспериментальных точек от y-координат графика функции была минимальной.

Мы не будем здесь производить подробное математическое описание метода наименьших квадратов. Достаточно того, что вы теперь знаете о существовании такого метода. Он очень широко используется в статистической обработке данных и встроен во многие математические пакеты программ. Важно понимать следующее: методом наименьших квадратов по данному набору экспериментальных точек можно построить любую (в том числе и из рассмотренных выше) функцию. А вот будет ли она нас удовлетворять, это уже другой вопрос — вопрос критерия соответствия. На рис. 3.5 изображены три функции, построенные методом наименьших квадратов по приведенным экспериментальным данным.

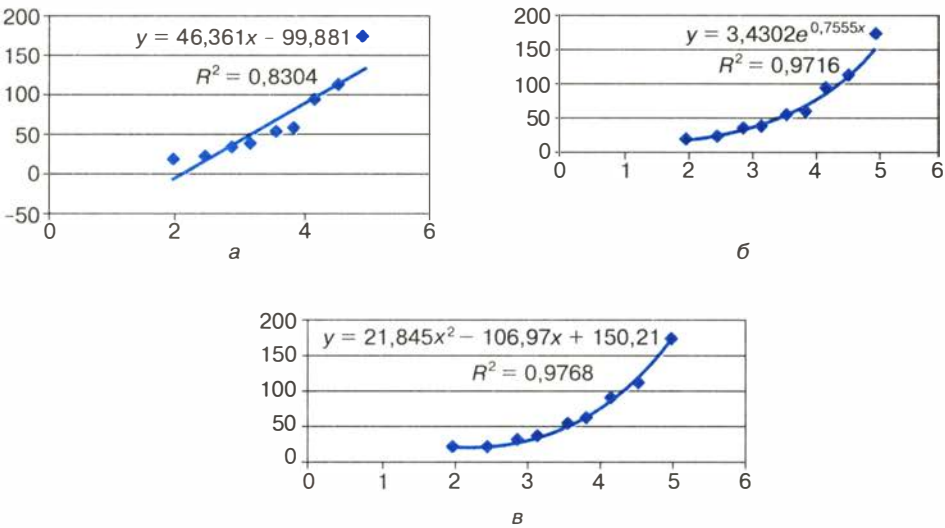

Рис. 3.5. Три функции, построенные по МНК

Эти рисунки получены с помощью табличного процессора Microsoft Excel. График регрессионной модели называется **трендом**. Английское слово *trend* можно перевести как «общее направление» или «тенденция».

Уже с первого взгляда хочется отбраковать вариант линейного тренда. График линейной функции — это прямая. Полученная по МНК прямая отражает факт роста заболеваемости от концентрации угарного газа, но по этому графику трудно что-либо сказать о характере этого роста. А вот квадратичный и экспоненциальный тренды правдоподобны. Теперь пора обратить внимание на надписи, присутствующие на графиках. Во-первых, это записанные в явном виде искомые функции — регрессионные модели:

линейная функция: $y = 46{,}361x - 99{,}881$;
экспоненциальная функция: $y = 3{,}4302\, e^{0{,}7555x}$;
квадратичная функция: $y = 21{,}845x^2 - 106{,}97x + 150{,}21$.

На графиках присутствует еще одна величина, полученная в результате построения трендов. Она обозначена как R^2. В статистике эта величина называется *коэффициентом детерминированности*. Именно она определяет, насколько удачной является полученная регрессионная модель. Коэффициент детерминированности всегда заключен в диапазоне от 0 до 1. Если он равен 1, то функция точно проходит через табличные значения, если 0, то выбранный вид регрессионной модели предельно неудачен. Чем R^2 ближе к 1, тем удачнее регрессионная модель.

Из трех выбранных моделей значение R^2 наименьшее у линейной. Значит, она самая неудачная (нам и так это было понятно). Значения же R^2 у двух других моделей достаточно близки (разница меньше 0,01). Если определить погрешность решения данной задачи как 0,01, по критерию R^2 эти модели нельзя разделить. Они одинаково удачны. Здесь могут вступить в силу качественные соображения. Например, если считать, что наиболее существенно влияние концентрации угарного газа проявляется при больших величинах, то, глядя на графики, предпочтение следует отдать квадратичной модели. Она лучше отражает резкий рост заболеваемости при больших концентрациях примеси.

Интересный факт: опыт показывает, что если человеку предложить на данной точечной диаграмме провести «на глаз» прямую так, чтобы точки были равномерно разбросаны вокруг нее, то он проведет линию, достаточно близкую к той, что дает МНК.

Прогнозирование по регрессионной модели

Мы получили регрессионную математическую модель и можем прогнозировать процесс путем вычислений. Теперь можно оценить уровень заболеваемости астмой не только для тех значений концентрации угарного газа, которые были получены путем измерений, но и для других значений. Это очень важно с практической точки зрения. Например, если в городе планируется построить завод, который будет выбрасывать в атмосферу угарный газ, то, рассчитав его возможную концентрацию, можно предсказать, как это отразится на заболеваемости астмой жителей города.

Существует два способа прогнозирования по регрессионной модели. Если прогноз производится в пределах экспериментальных значений независимой переменной (в нашем случае это концентрация угарного газа C), то это называется *восстановлением значения.*

Прогнозирование за пределами экспериментальных данных называется *экстраполяцией.*

Имея регрессионную модель, легко прогнозировать, производя расчеты с помощью электронных таблиц. Выберем для нашего примера в качестве наиболее подходящей квадратичную зависимость. Построим следующую электронную таблицу:

	A	B
1	*Концентрация угарного газа (мг/куб. м)*	*Число больных астмой на 1 тыс. жителей*
2		=21,845*A2*A2–106,97*A2+150,21

Подставляя в ячейку A2 значение концентрации угарного газа, в ячейке B2 будем получать прогноз заболеваемости. Вот пример восстановления значения:

	A	B
1	*Концентрация угарного газа (мг/куб. м)*	*Число больных астмой на 1 тыс. жителей*
2	3	25

Заметим, что число, получаемое по формуле в ячейке B2, на самом деле является дробным. Однако не имеет смысла считать число людей, даже среднее, в дробных величинах. Дробная часть удалена — в формате вывода числа указано 0 цифр после запятой.

Экстраполяционный прогноз выполняется аналогично.

Табличный процессор дает возможность производить экстраполяцию графическим способом, продолжая тренд за пределы экспериментальных данных. Как это выглядит при использовании квадратичного тренда для $C = 7$, показано на рис. 3.6.

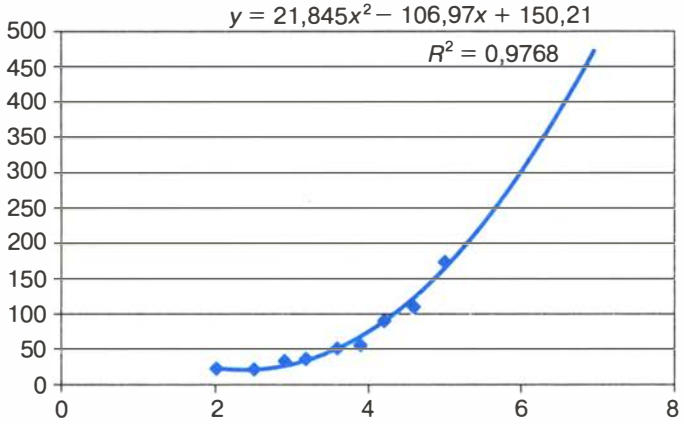

Рис. 3.6. Квадратичный тренд с экстраполяцией

В ряде случаев с экстраполяцией надо быть осторожным. Применимость всякой регрессионной модели ограничена, особенно за пределами экспериментальной области. В нашем примере при экстраполяции не следует далеко уходить от величины 5 мг/м^3. Вполне возможно, что далее характер зависимости существенно меняется. Слишком сложной является система «экология — здоровье человека», в ней много различных факторов, которые связаны друг с другом. Полученная регрессионная функция является всего лишь моделью, экспериментально подтвержденной в диапазоне концентраций от 2 до 5 мг/м^3. Что будет вдали от этой области, мы не знаем. Всякая экстраполяция держится на гипотезе: «предположим, что за пределами экспериментальной области закономерность сохраняется». А если не сохраняется?

Квадратичная модель в данном примере в области малых значений концентрации, близких к 0, вообще не годится. Экстраполируя ее на $C = 0$ мг/м^3, получим 150 человек больных, т. е. больше, чем при 4 мг/м^3. Очевидно, это нелепость. В области малых значений C лучше работает экспоненциальная модель. Кстати, это довольно типичная ситуация: разным областям данных могут лучше соответствовать разные модели.

Система основных понятий

<table>
<tr><th colspan="3">Модели статистического прогнозирования</th></tr>
<tr><td colspan="3">Статистика: наука о сборе, измерении и анализе массовых количественных данных</td></tr>
<tr><td colspan="3">Статистические данные</td></tr>
<tr><td colspan="2">Приближенный характер</td><td>Требуют многократных измерений</td></tr>
<tr><td colspan="3">Регрессионная модель</td></tr>
<tr><td>Описывает зависимость между количественными характеристиками сложных систем</td><td>Вид регрессионной функции определяется подбором по экспериментальным данным</td><td>Может использоваться для прогнозирования</td></tr>
<tr><td colspan="3">Метод наименьших квадратов</td></tr>
<tr><td>Используется для вычисления параметров регрессионной модели</td><td>Вид регрессионной модели задает пользователь</td><td>Содержится в математическом арсенале электронных таблиц</td></tr>
</table>

Вопросы и задания

1. а) Что такое статистика?
   б) Являются ли результаты статистических расчетов точными?
   в) Что такое регрессионная модель?
2. Какие из следующих величин можно назвать статистическими: температура вашего тела в данный момент; средняя температура в вашем регионе за последний месяц; максимальная скорость, развиваемая данной моделью автомобиля; среднее число осадков, выпадающих в вашем регионе в течение года?
3. а) Для чего используется метод наименьших квадратов?
   б) Что такое тренд?
   в) Как располагается линия тренда, построенная по МНК, относительно экспериментальных точек?
   г) Может ли тренд, построенный по МНК, пройти выше всех экспериментальных точек?
4. а) В чем смысл параметра R^2? Какие значения он принимает?
   б) Какое значение примет параметр R^2, если тренд точно проходит через экспериментальные точки?
5. По данным из следующей таблицы постройте с помощью Excel линейную, квадратичную, экспоненциальную и логарифмическую регрессионные модели. Определите параметры, выберите лучшую модель.

x	2	4	6	8	10	12	14	16	18	20	22	24	26	28
y	44	32	35	40	30	27	21	25	20	23	18	19	20	16

6. а) Что подразумевается под восстановлением значения по регрессионной модели ?
   б) Что такое экстраполяция?
7. Соберите данные о средней дневной температуре в вашем городе за последнюю неделю (10 дней, 20 дней). Оцените (хотя бы на глаз), годится ли использование линейного тренда для описания характера изменения температуры со временем. Попробуйте путем графической экстраполяции предсказать температуру через 2–5 дней.
8. Придумайте свои примеры практических задач, для которых имело бы смысл выполнение восстановления значений и экстраполяционных расчетов.

§ 19 Моделирование корреляционных зависимостей

Регрессионные математические модели строятся в тех случаях, когда известно, что зависимость между двумя факторами существует и требуется получить ее математическое описание. А сейчас мы рассмотрим задачи другого рода. Пусть важной характеристикой некоторой сложной системы является фактор A. На него могут оказывать влияние одновременно многие другие факторы: B, C, D и т. д. Мы рассмотрим два типа задач.

1) Оказывает ли фактор B какое-либо заметное регулярное влияние на фактор A?
2) Какие из факторов B, C, D и т. д. оказывают наибольшее влияние на фактор A?

В качестве примера сложной системы будем рассматривать школу. Пусть для первого типа задач фактором A является средняя успеваемость учащихся школы, фактором B — финансовые расходы школы на хозяйственные нужды: ремонт здания, обновление мебели, эстетическое оформление помещения и т. п. Здесь влияние фактора B на фактор A не очевидно. Наверное, гораздо сильнее на успеваемость влияют другие причины: уровень квалификации учителей, контингент учащихся, уровень технических средств обучения и др.

Специалисты по статистике знают, что для того, чтобы выявить зависимость от какого-то определенного фактора, нужно максимально исключить влияние других факторов. Проще говоря,

собирая информацию из разных школ, нужно выбирать такие школы, в которых приблизительно одинаковый контингент учеников, квалификация учителей и пр., но хозяйственные расходы разные (у одних школ могут быть богатые спонсоры, у других — нет).

Итак, пусть хозяйственные расходы школы выражаются количеством рублей, отнесенных к числу учеников в школе (руб./чел.), потраченных за определенный период времени (например, за последние 5 лет). Успеваемость же пусть оценивается средним баллом учеников школы по результатам окончания последнего учебного года. Еще раз обращаем ваше внимание на то, что в статистических расчетах обычно используются относительные и усредненные величины.

Итоги сбора данных по 20 школам, введенные в электронную таблицу, представлены на рис. 3.7. На рис. 3.8 приведена точечная диаграмма, построенная по этим данным.

A	B	C
№ п/п	Затраты (руб./чел.)	Успеваемость (средний балл)
1	50	3,81
2	345	4,13
3	79	4,30
4	100	3,96
5	203	3,87
6	420	4,33
7	210	4
8	137	4,21
9	463	4,4
10	231	3,99
11	134	3,9
12	100	4,07
13	294	4,15
14	396	4,1
15	77	3,76
16	480	4,25
17	450	3,88
18	496	4,50
19	102	4,12
20	150	4,32

Рис. 3.7. Статистические данные

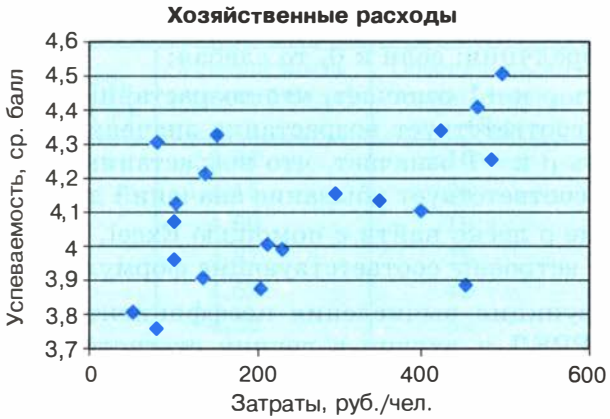

Рис. 3.8. Точечная диаграмма

Значения обеих величин: финансовых затрат и успеваемости учеников — имеют значительный разброс и, на первый взгляд, взаимосвязи между ними не видно. Однако она вполне может существовать.

Зависимости между величинами, каждая из которых подвергается не контролируемому полностью разбросу, называются **корреляционными зависимостями**.

Раздел математической статистики, который исследует такие зависимости, называется **корреляционным анализом**. Корреляционный анализ изучает усредненный закон поведения каждой из величин в зависимости от значений другой величины, а также меру такой зависимости.

Оценку корреляции величин начинают с высказывания гипотезы о возможном характере зависимости между их значениями. Чаще всего допускают наличие линейной зависимости. В таком случае мерой корреляционной зависимости является величина, которая называется **коэффициентом корреляции**. Как и прежде, мы не будем писать формулы, по которым этот коэффициент вычисляется; их написать нетрудно, гораздо труднее понять, почему они именно такие. На данном этапе достаточно знать следующее:

- коэффициент корреляции (обычно обозначаемый греческой буквой ρ) есть число из диапазона от -1 до $+1$;

- если это число по модулю близко к 1, то имеет место сильная корреляция; если к 0, то слабая;
- близость ρ к +1 означает, что возрастанию значений одного набора соответствует возрастание значений другого набора, близость ρ к –1 означает, что возрастанию значений одного набора соответствует убывание значений другого набора;
- значение ρ легко найти с помощью Excel, так как в эту программу встроены соответствующие формулы.

В Excel функция вычисления коэффициента корреляции называется КОРРЕЛ и входит в группу статистических функций. Покажем, как ею воспользоваться. На том же листе Excel, где находится таблица, представленная на рис. 3.7, надо установить курсор на любую свободную ячейку и запустить функцию КОРРЕЛ. Она запросит два диапазона значений. Укажем, соответственно, B2:B21 и C2:C21. После их ввода будет выведен ответ: $\rho = 0{,}500273843$. Эта величина говорит о среднем уровне корреляции.

Наличие зависимости между хозяйственными затратами школы и успеваемостью нетрудно понять. Ученики с удовольствием ходят в чистую, красивую, уютную школу, чувствуют там себя, как дома, и поэтому лучше учатся.

В следующем примере проводится исследование по определению зависимости успеваемости учащихся старших классов от двух факторов: обеспеченности школьной библиотеки учебниками и оснащения школы компьютерами. И та, и другая характеристика количественно выражается в процентах от нормы. Нормой обеспеченности учебниками является их полный комплект, т. е. такое количество, когда каждому ученику выдаются из библиотеки все нужные ему для учебы книги. Нормой оснащения компьютерами будем считать такое их количество, при котором на каждых четырех старшеклассников в школе приходится один компьютер. Предполагается, что компьютерами ученики пользуются не только на информатике, но и на других уроках, а также во внеурочное время.

В таблице, изображенной на рис. 3.9, приведены результаты измерения обоих факторов в 11 разных школах. Напомним, что влияние каждого фактора исследуется независимо от других (т. е. влияние других существенных факторов должно быть приблизительно одинаковым).

Обеспечение учебного процесса				
№	Обеспеченность учебниками (%)	Успеваемость (средний балл)	Обеспеченность компьютерами (%)	Успеваемость (средний балл)
1	50	3,81	10	3,98
2	78	4,15	25	4,01
3	94	4,69	19	4,34
4	65	4,37	78	4,41
5	99	4,53	45	3,94
6	87	4,23	32	3,62
7	100	4,73	90	4,6
8	63	3,69	21	4,24
9	79	4,08	34	4,36
10	94	4,2	45	3,99
11	93	4,32	67	4,5
	$\rho = 0{,}780931$		$\rho = 0{,}572465$	

Рис. 3.9. Сравнение двух корреляционных зависимостей

Для обеих зависимостей получены коэффициенты линейной корреляции. Как видно из таблицы, корреляция между обеспеченностью учебниками и успеваемостью сильнее, чем корреляция между компьютерным обеспечением и успеваемостью (хотя и тот, и другой коэффициенты корреляции не очень большие). Отсюда можно сделать вывод, что пока еще книга остается более значительным источником знаний, чем компьютер.

Система основных понятий

Корреляционные зависимости	
Это зависимости между величинами, каждая из которых подвергается неконтролируемому разбросу	
Корреляционный анализ дает возможность:	
определить, оказывает ли один фактор существенное влияние на другой фактор	выбрать из нескольких факторов наиболее существенный
Коэффициент корреляции ρ: количественная мера корреляции	
ρ по модулю близко к единице — сильная корреляция	ρ близко к нулю — слабая корреляция
Расчет ρ возможен в Microsoft Excel с помощью функции КОРРЕЛ	

Вопросы и задания

1. а) Что такое корреляционная зависимость?
   б) Что такое корреляционный анализ?
   в) Какие типы задач можно решать с помощью корреляционного анализа?
   г) Какая величина является количественной мерой корреляции? Какие значения она может принимать?
2. С помощью какого средства табличного процессора Excel можно вычислить коэффициент корреляции?
3. а) Для данных из таблицы, представленной на рис. 3.9, постройте две линейные регрессионные модели.
   б) Для этих же данных вычислите коэффициенты корреляции. Сравните с приведенными на рис. 3.9 результатами.

§ 20
Модели оптимального планирования

Проблема, к обсуждению которой мы теперь переходим, называется **оптимальным планированием.** Объектами планирования могут быть самые разные системы: деятельность отдельного предприятия, отрасли промышленности или сельского хозяйства, региона, наконец государства. Постановка задачи планирования выглядит следующим образом:

- имеются некоторые плановые показатели: X, Y, и др.;
- имеются некоторые ресурсы: $R1$, $R2$ и др., за счет которых эти плановые показатели могут быть достигнуты. Эти ресурсы практически всегда ограничены;
- имеется определенная стратегическая цель, зависящая от значений X, Y и др. плановых показателей, на которую следует ориентировать планирование.

Нужно определить значение плановых показателей с учетом ограниченности ресурсов при условии достижения стратегической цели. Это и будет оптимальным планом.

Приведем примеры. Пусть объектом планирования является детский сад. Ограничимся лишь двумя плановыми показателями: количеством детей и количеством воспитателей. Основными ресурсами деятельности детского сада являются объем финансирования и площади помещения. А каковы стратегические цели?

Естественно, одной из них является сохранение и укрепление здоровья детей. Количественной мерой такой цели является минимизация заболеваемости воспитанников детского сада.

Другой пример: планирование экономической деятельности государства. Безусловно, это слишком сложная задача для того, чтобы нам с ней полностью разобраться. Плановых показателей очень много: это производство различных видов промышленной и сельскохозяйственной продукции, подготовка специалистов, выработка электроэнергии, размер зарплаты работников бюджетной сферы и многое другое. К ресурсам относятся: количество работоспособного населения, бюджет государства, природные ресурсы, энергетика, возможности транспортных систем и пр. Как вы понимаете, каждый из этих видов ресурсов ограничен. Кроме того, важнейшим ресурсом является время, отведенное на выполнение плана. Вопрос о стратегических целях довольно сложный. У государства их много, но в разные периоды истории приоритеты целей могут меняться. Например, в военное время главной целью является максимальная обороноспособность, военная мощь страны. В мирное время в современном цивилизованном государстве приоритетной целью должно быть достижение максимального уровня жизни населения.

Если мы хотим использовать компьютер для решения задачи оптимального планирования, то нам снова нужно построить математическую модель. Следовательно, всё, о чем говорилось в примерах, должно быть переведено на язык чисел, формул, уравнений и других средств математики. В полном объеме для реальных систем эта задача очень сложная. Как и раньше, мы пойдем по пути упрощения. Рассмотрим очень простой пример, из которого вы получите представление об одном из подходов к решению задачи оптимального планирования.

Пример. Школьный кондитерский цех готовит пирожки и пирожные. В силу ограниченности емкости склада за день можно приготовить в совокупности не более 700 штук изделий. Рабочий день в кондитерском цехе длится 8 часов. Производство пирожных более трудоемко, поэтому если выпускать только их, за день можно произвести не более 250 штук, пирожков же можно произвести 1000 штук (если при этом не выпускать пирожных). Стоимость пирожного вдвое выше, чем стоимость пирожка. Требуется составить такой дневной план производства, чтобы обеспечить наибольшую выручку кондитерского цеха.

Разумеется, это чисто учебный пример. Вряд ли существует такой кондитерский цех, который выпускает всего два вида продукции, да и наибольшая выручка — не единственная цель его работы. Но зато математически формулировка задачи будет простой. Давайте ее выработаем.

Плановыми показателями являются:

- x — дневной план выпуска пирожков;
- y — дневной план выпуска пирожных.

Что в этом примере можно назвать ресурсами производства? Из того, о чем говорится в условии задачи, это:

- длительность рабочего дня — 8 часов;
- вместимость складского помещения — 700 мест.

Предполагается для простоты, что другие ресурсы (сырье, электроэнергия и пр.) не ограничены. Формализацию цели (достижение максимальной выручки цеха) мы обсудим позже.

Получим соотношения, следующие из условий ограниченности времени работы цеха и вместимости склада, т. е. суммарного числа изделий.

Из постановки задачи следует, что на изготовление одного пирожного затрачивается в 4 раза больше времени, чем на выпечку одного пирожка. Если обозначить время изготовления пирожка как t мин, то время изготовления пирожного будет равно $4t$ мин. Значит, суммарное время на изготовление x пирожков и y пирожных равно

$$tx + 4ty = (x + 4y)t.$$

Но это время не может быть больше длительности рабочего дня. Отсюда следует неравенство:

$$(x + 4y)t \le 8 \cdot 60,$$

или

$$(x + 4y)t \le 480.$$

Легко посчитать t — время изготовления одного пирожка. Поскольку за рабочий день их может быть изготовлено 1000 штук, на один пирожок тратится $480/1000 = 0{,}48$ мин. Подставляя это значение в неравенство, получим:

$$(x + 4y) \cdot 0{,}48 \le 480.$$

Отсюда

$$x + 4y \le 1000.$$

Ограничение на общее число изделий дает совершенно очевидное неравенство:

$$x + y \le 700.$$

К двум полученным неравенствам следует добавить условия положительности значений величин x и y (не может быть отрицательного числа пирожков и пирожных). В итоге получим систему неравенств:

$$\begin{cases} x + 4y \le 1000; \\ x + y \le 700; \\ x \ge 0; \\ y \ge 0. \end{cases} \qquad (1)$$

А теперь перейдем к формализации стратегической цели: получению максимальной выручки. Выручка — это стоимость всей проданной продукции. Пусть цена одного пирожка — r рублей. По условию задачи, цена пирожного в два раза больше, т. е. $2r$ рублей. Отсюда стоимость всей произведенной за день продукции равна

$$rx + 2ry = r(x + 2y).$$

Целью производства является получение максимальной выручки. Будем рассматривать записанное выражение как функцию от x, y:

$$F(x, y) = r(x + 2y).$$

Она называется **целевой функцией**.

Поскольку значение r — константа, максимальное значение $F(x, y)$ будет достигнуто при максимальной величине выражения $(x + 2y)$. Поэтому в качестве целевой функции можно принять

$$f(x, y) = x + 2y. \qquad (2)$$

Следовательно, получение оптимального плана свелось к следующей математической задаче:

Требуется найти значения плановых показателей x и y, удовлетворяющих данной системе неравенств (1) *и придающих максимальное значение целевой функции* (2).

Итак, математическая модель задачи оптимального планирования для школьного кондитерского цеха построена.

Теперь следующий вопрос: как решить эту задачу? Вы уже догадываетесь, что решать ее за нас будет компьютер с помощью табличного процессора Excel. А мы обсудим лишь подход к решению, не вникая в подробности метода.

Математическая дисциплина, которая посвящена решению таких задач, называется **математическим программированием**. А поскольку в целевую функцию $f(x, y)$ величины x и y входят линейно (т. е. в первой степени), наша задача относится к разделу этой науки, который называется *линейным программированием*.

Система написанных выше неравенств представляется на координатной плоскости четырехугольником, ограниченным четырьмя прямыми, соответствующими линейным уравнениям:

$$x + 4y = 1000,$$
$$x + y = 700,$$
$$x = 0 \quad (\text{ось } Y),$$
$$y = 0 \quad (\text{ось } X).$$

На рис. 3.10 эта область представляет собой четырехугольник $ABCD$ и выделена заливкой. Любая точка четырехугольника является решением системы неравенств (1). Например, $x = 200$, $y = 100$. Этой точке соответствует значение целевой функции

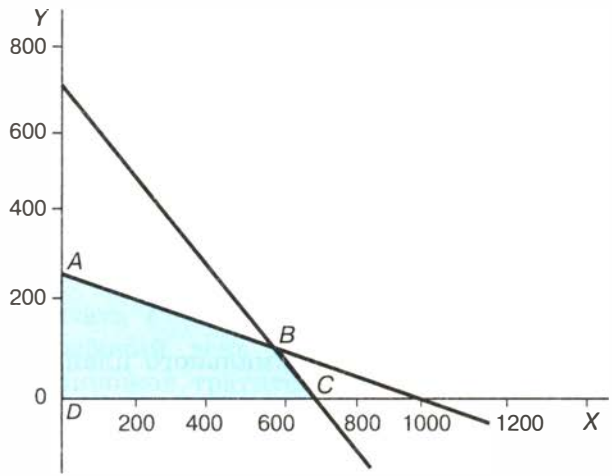

Рис. 3.10. Область поиска оптимального плана

$f(200, 100) = 400$. А другой точке ($x = 600$, $y = 50$) соответствует $f(600, 50) = 700$. Но, очевидно, искомым решением является та точка области $ABCD$, в которой целевая функция максимальна. Нахождение этой точки производится с помощью методов линейного программирования.

В математическом арсенале Excel имеется средство **Поиск решения**. Как решать данную задачу с помощью этого средства, вы узнаете из компьютерного практикума.

В результате решения задачи получается следующий оптимальный план дневного производства кондитерского цеха: нужно выпускать 600 пирожков и 100 пирожных. Эти плановые показатели соответствуют координатам точки B на рис. 3.10. В этой точке значение целевой функции $f(600, 100) = 800$. Если один пирожок стоит 5 рублей, то полученная выручка составит 4000 рублей.

Система основных понятий

<table>
<tr><th colspan="3">Модели оптимального планирования</th></tr>
<tr><td colspan="3">Оптимальное планирование — определение значений плановых показателей с учетом ограниченности ресурсов при условии достижения заданной цели</td></tr>
<tr><td colspan="3">Ограниченность ресурсов описывается:</td></tr>
<tr><td>системой неравенств</td><td>системой равенств</td><td>смешанной системой</td></tr>
<tr><td colspan="3">Цель описывается функцией, для которой требуется</td></tr>
<tr><td colspan="2">найти минимум</td><td>найти максимум</td></tr>
<tr><td colspan="3">Microsoft Excel имеет специальное средство Поиск решения для решения задач оптимального планирования</td></tr>
</table>

Вопросы и задания

1. а) В чем состоит задача оптимального планирования?
   б) Что такое плановые показатели, ресурсы, стратегическая цель? Приведите примеры.
2. а) Попробуйте сформулировать содержание оптимального планирования для своей учебной деятельности.

б) Что такое математическое программирование, линейное программирование?

3. а) Сформулируйте задачу оптимального планирования для школьного кондитерского цеха, в котором выпускается три вида продукции: пирожки, пирожные и коржики.
б) Внесите изменение в постановку задачи оптимального планирования из этого параграфа для двух видов продукции с учетом еще одного ограничения: число пирожных должно быть не меньше числа пирожков. На координатной плоскости постройте область поиска решения.

ЭОР к главе 3 на сайте ФЦИОР (http://fcior.edu.ru)

- Назначение и виды информационных моделей
- Построение информационных моделей ИС
- Формализация задач из различных предметных областей. Формирование требований к ИС