МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Пензенский государственный университет
архитектуры и строительства

В.А. Смирнов

ПРИКЛАДНАЯ СТАТИСТИКА В ПАКЕТЕ АНАЛИЗА MS EXCEL

Рекомендовано редсоветом университета в качестве учебного пособия для студентов, обучающихся по специальности 200500 «Метрология, стандартизация и сертификация»

Пенза 2008

УДК 004.2
ББК 22.151
С 50

Рецензент – зав. кафедрой стандартизации, сертификации и контроля качества, профессор В.И. Логанина (ПГУАС).

Смирнов, В.А.
С 50 Прикладная статистика в пакете анализа MS Excel [текст]: учебное пособие / В.А. Смирнов. – Пенза: ПГУАС, 2008. - 88 с.

Излагаются теоретические основы ряда методов прикладной статистики. На конкретных примерах рассматриваются постановка и методы решения статистических задач. Приводятся практические рекомендации по использованию пакета анализа MS Excel, предназначенного для статистической обработки данных.

Учебное пособие подготовлено на кафедре математики и математического моделирования и предназначено для студентов вузов, обучающихся по направлению 200500 «Метрология, стандартизация и сертификация».

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Настоящее учебное пособие является частью курса лекций «Программные статистические комплексы».

Рассматриваются методы описательной статистики, аппарат проверки статистических гипотез, методы регрессионного анализа и основы математической теории эксперимента. Излагаются некоторые вопросы, связанные с построением и анализом экспериментально-статистических моделей, линейных по параметрам.

Все рассмотренные методы сопровождаются примерами решения соответствующих задач в пакете анализа MS Excel. В приложении приведен пример построения и анализа квадратичной двухфакторной модели с использованием программного комплекса «Градиент», разработанного в ПГУАС.

Учебное пособие подготовлено на кафедре математики и математического моделирования ПГУАС на основании лекций и практических занятий по дисциплинам «Математика» и «Программные статистические комплексы». Пособие предназначено для студентов вузов, обучающихся по направлению 200500 «Метрология, стандартизация и сертификация», однако может оказаться полезным и для студентов других специальностей.

ВВЕДЕНИЕ

Анализ эмпирической информации и получение обоснованных выводов невозможны без использования методов математической статистики.

Целесообразность применения программных средств, реализующих методы прикладной статистики – *программных статистических комплексов*, или *статистических пакетов*, в основном определяется двумя обстоятельствами.

Во-первых, объем подлежащей анализу информации достаточно велик. Необходимость работы с большими массивами данных затрудняет вычисления с использованием простейших средств.

Во-вторых, для методов математической статистики характерно использование большого числа специальных функций, нахождение значений которых затруднительно; безмашинные методы требуют работы с громоздкими таблицами.

Широкому внедрению машинных методов статистического анализа способствовало распространение персональных компьютеров и появление соответствующих программных средств. Среди последних особое место занимает группа пакетов статистического анализа, *входящихявсоставяпрограммныхяпродуктовясходногояназначения* (табличных процессоров, систем управления базами данных, систем визуализации). В состав пакетов этой группы входят средства реализации методов описательной статистики, методов проверки статистических гипотез, методов регрессионного анализа.

По сравнению со специализированными и универсальными программными статистическими комплексами пакеты анализа отличает *доступность*; в частности, пакеты анализа входят в состав табличного процессора MS Excel, а также табличных процессоров свободно распространяемых пакетов OpenOffice, KOffice и GNOME Office.

Большинство программных продуктов, включающих пакеты анализа, содержит также и встроенный *командныйяязык* – интерпретируемый алгоритмический язык высокого уровня, на котором могут быть описаны нестандартные задачи.

1. Методы описательной статистики

Методы предназначены для первичного анализа большой выборки значений одного признака.

Пусть из генеральной совокупности X извлечена выборка

$$\{(x_i, n_i)\}, \ i = \overline{1, n}, \tag{1.1}$$

где n − объем выборки, n_i − число появлений значения x_i.

Наблюдаемые значения называют *вариантами*. Число n_i появлений значения x_i называют *частотой*, а частное n_i/n от деления частоты на объем выборки − *относительной частотой*. Последовательность вариант и соответствующих им частот, упорядоченная в возрастающем порядке, называется *дискретным вариационным рядом*.

Если объем выборки значителен, то дискретный вариационный ряд теряет наглядность. В этом случае выполняют группировку данных − построение *непрерывного вариационного ряда*.

При выполнении группировки весь диапазон $[x_{\min}; x_{\max}]$ изменения величины x делится на несколько интервалов − *разрядов*, число которых выбирают по *правилу Стерджеса*:

$$l = 1 + 3{,}3 \lg n. \tag{1.2}$$

Частоты, соответствующие каждому разряду, находятся как суммы частот всех вариант, попавших в этот разряд (если в исходной выборке каждая варианта встречается только один раз, то частота находится как количество вариант, попавших в интервал).

Для графического представления непрерывного вариационного ряда выполняют построение *гистограммы* − ступенчатой фигуры, состоящей из прямоугольников, основания которых построены на соответствующих разрядах, а высоты h_j равны частному от деления относительной частоты на длину разряда:

$$h_j = \frac{n_j}{n(x_{j+1} - x_j)}, \ j = \overline{1, l}. \tag{1.3}$$

Гистограмма позволяет сделать предварительное суждение о плотности распределении генеральной совокупности.

Статистическими оценками называют функции от наблюдаемых значений. *Точечными оценками* называют оценки, выражаемые одним числом.

Положение «центра» распределения может быть охарактеризовано тремя различными точечными оценками − оценкой медианы, оценкой моды и оценкой математического ожидания.

Если при построении дискретного вариационного ряда варианту с частотой m записать ровно m раз, то в качестве *оценки медианы* следует взять значение, соответствующее центру ряда:

$$Me = \begin{cases} x_{(n+1)/2}, & n = 2k+1 \\ \dfrac{x_{n/2} + x_{n/2+1}}{2}, & n = 2k \end{cases} . \qquad (1.4)$$

Оценку моды обычно находят графически. Для этого на гистограмме находят прямоугольник с наибольшей высотой и проводят из противоположных вершин его верхнего основания два отрезка к противоположным вершинам верхних оснований соседних прямоугольников. В качестве оценки моды принимается абсцисса точки пересечения этих отрезков.

п ценкойя математическогоя ожидания является *выборочноея среднее* − среднее арифметическое вариант

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i . \qquad (1.5)$$

Для характеристики «рассеяния» значений около «центра» используют оценки дисперсии, среднего квадратичного и среднего абсолютного отклонения.

В качестве несмещенной *оценкиядисперсии* используют величину

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 . \qquad (1.6)$$

Оценка *стандартного* (*среднегоя квадратичного*) *отклонения* связана с оценкой дисперсии соотношением

$$s = \sqrt{s^2} . \qquad (1.7)$$

Стандартнаяя ошибка оценки математического ожидания вычисляется как частное от деления стандартного отклонения на квадратный корень из объема выборки (как корень из частного от деления дисперсии на объем выборки).*я*

п ценкаясреднегояабсолютногояотклонения равна

$$Adev = \frac{1}{n}\sum_{i=1}^{n} \left|x_i - \bar{x}\right| . \qquad (1.8)$$

Характеристиками рассеяния вариант также являются *нижняя* $x_{1/4}$ и *верхняя* $x_{3/4}$ *квартили* − значения, для которых число вариант, удовлетворяющих неравенствам $x_i < x_{1/4}$ и $x_i < x_{3/4}$, составляет 25% и 75%, соответственно.

Оценки моментов третьего и четвертого порядков и связанные с ними безразмерные величины – оценки асимметрии и эксцесса – используются реже. Оценка асимметрии

$$Skew = \frac{1}{ns^3}\sum_{i=1}^{n}\left(x_i - \vec{x}\right)^3 \qquad (1.9)$$

характеризует «скос» распределения относительно его «центра» в положительном или отрицательном направлениях, соответственно.

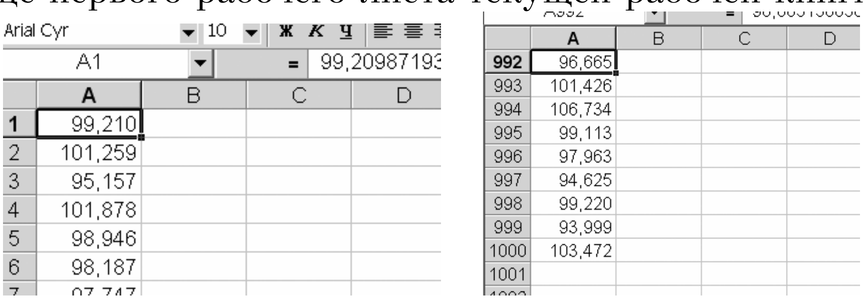
Оценка эксцесса

$$Kurt = \left(\frac{1}{ns^4}\sum_{i=1}^{n}\left(x_i - \vec{x}\right)^4\right) - 3 \qquad (1.10)$$

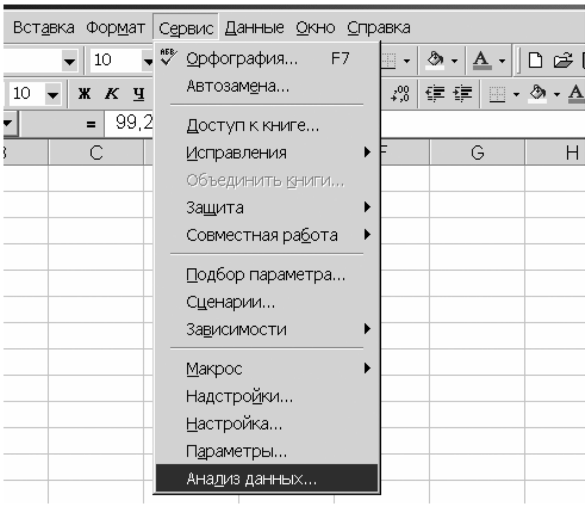
характеризует «островершинность» (при $Kurt > 0$) или «плосковершинность» (при $Kurt < 0$) распределения по сравнению с нормальным.

1.1. Решение задач описательной статистики с помощью пакета анализа MS Excel

Пусть выборка, содержащая 1000 вариант, расположена в первом столбце первого рабочего листа текущей рабочей книги:

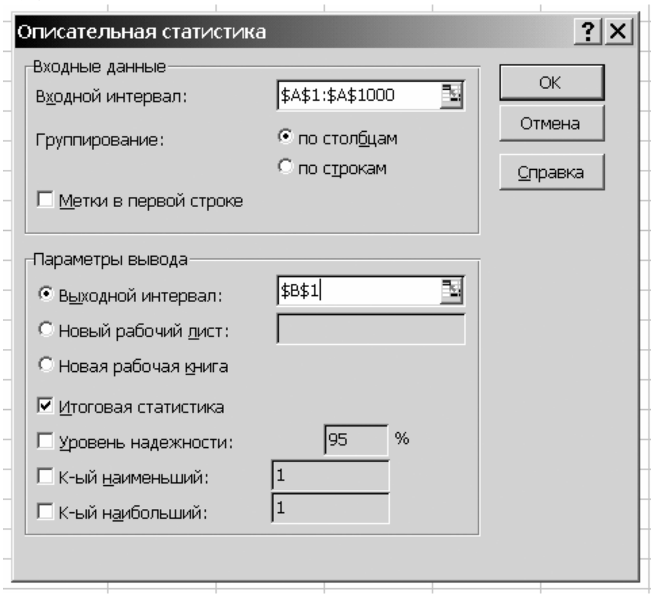


Для нахождения точечных оценок распределения следует из меню Сервис выбрать пункт Анализ данных:



Если указанный пункт меню недоступен, то необходимо установить пакет анализа (выбор Сервис – Надстройки; в диалоговом окне установить флажок Пакет анализа).

В списке Инструменты анализа следует выбрать пункт Описательная статистика. В диалоговом окне Описательная статистика



необходимо указать диапазон рабочего листа, содержащий выборку; в данном примере – `$A$1:$A$1000`. В качестве выходного интервала достаточно указать первую ячейку второго столбца – `$B$1`. Дополнительно следует установить флажок Итоговая статистика, после чего нажать Enter. Результаты анализа будут помещены во второй столбец:

	A	B	C
1	99,210	*Столбец1*	
2	101,259		
3	95,157	Среднее	99,91778
4	101,878	Стандартная ошибка	0,093233
5	98,946	Медиана	99,85605
6	98,187	Мода	97,38525
7	97,747	Стандартное отклонение	2,948278
8	100,798	Дисперсия выборки	8,692346
9	99,088	Эксцесс	-0,17834
10	104,529	Асимметричность	0,074057
11	98,301	Интервал	17,02596
12	101,805	Минимум	91,45482
13	102,784	Максимум	108,4808
14	95,570	Сумма	99917,78
15	102,903	Счет	1000
16	97,288		

Указать диапазон, содержащий выборку, можно следующим образом: после перевода фокуса ввода на поле Входной интервал щелкнуть на первой ячейке диапазона (`$A$1`); затем, удерживая клавиши Shift и Control, нажать PageDown; при этом диапазон будет расширен до последней заполненной ячейки (`$A$1000`).

Пакет анализа MS Excel содержит встроенные средства построения непрерывного вариационного ряда и гистограммы, однако эти средства функционируют не вполне корректно. Поэтому часть данных для анализа следует подготовить отдельно.

Найдем границы разрядов. Интервал изменения вариант – от `91,5` до `108,5` – уже известен. В качестве левой границы первого разряда выберем `90`, в качестве правой границы последнего `110`.

Так как

$$l = 1 + 3{,}3\lg 1000 = 10{,}9,$$

то число разрядов можно взять равным 10, а длина каждого разряда равна

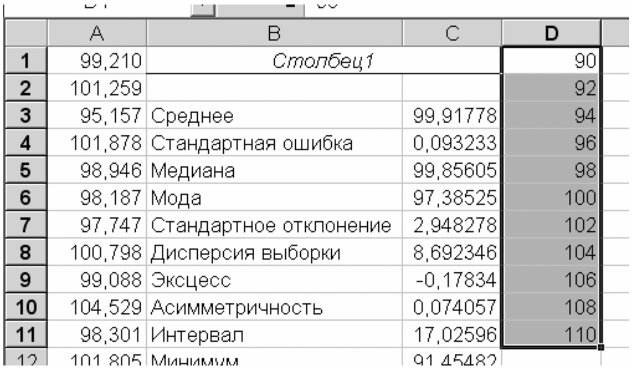
$$\frac{110-90}{10} = 2.$$

Вычисление границ удобно выполнять с использованием автозаполнения. После двойного щелчка на ячейке `D1` вводится `90`; нажатие на Enter переводит на ячейку `D2`. В эту ячейку следует ввести число `92`. Затем следует выделить ячейки `D1` и `D2` (щелчок на `D1`, нажать и удерживать `Shift`, щелчок на `D2`), подвести курсор к маркеру автозаполнения (черный квадрат в правом нижнем углу ячейки `D2`):

D1 = 90

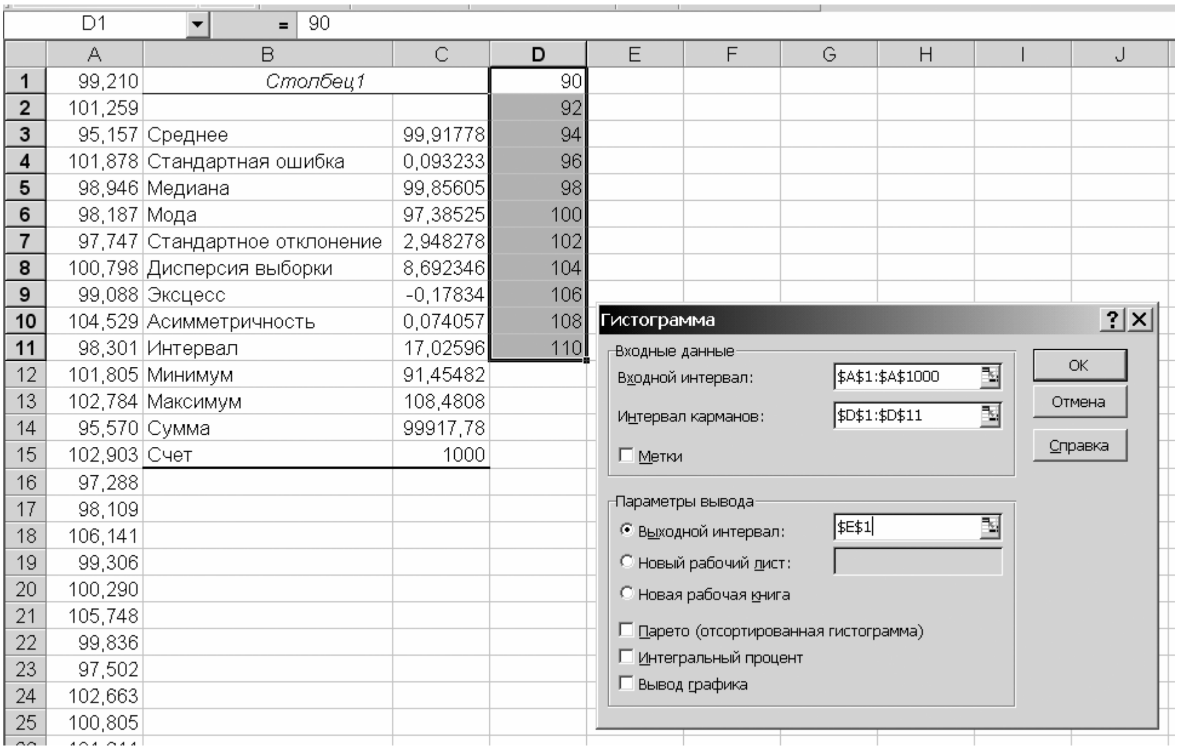
	A	B	C	D
1	99,210	*Столбец1*		90
2	101,259			92

и, удерживая левую клавишу мыши, перевести маркер до ячейки `D11`:



.

После этого из меню **Сервис** вновь следует выбрать **Анализ данных**, и в списке инструментов анализа выбрать пункт **Гистограмма**. Как и ранее, входным интервалом вновь будет диапазон `$A$1:$A$1000`. Интервал, содержащий границы разрядов, указывается в поле **Интервал карманов** (в данном примере – `$D$1:$D$11`). В качестве выходного интервала достаточно указать первую ячейку пятого столбца – `$E$1`:



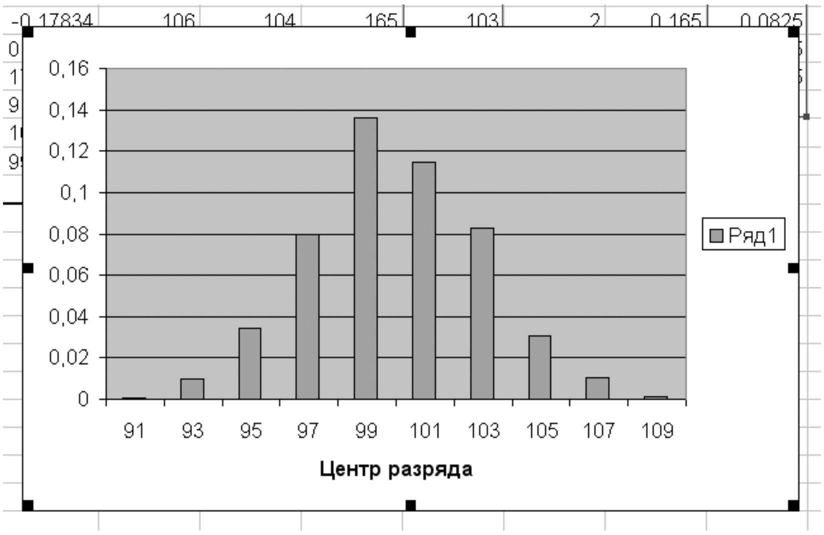
Частоты, соответствующие каждому разряду, помещаются в ячейки `F3:F12`:

	D	E	F
	90	*Карман*	*Частота*
	92	90	0
78	94	92	1
33	96	94	20
05	98	96	69
25	100	98	160
78	102	100	272
46	104	102	229
34	106	104	165
57	108	106	61
96	110	108	21
32		110	2
08		Еще	0
78			

Перед построением гистограммы следует:

- вычислить значения, соответствующие центру каждого разряда – в ячейку `G3` вводится
  `=(E3+E2)/2`
  нажатие Enter, маркер автозаполнения переводится от ячейки `G3` до ячейки `G12`;
- вычислить длины каждого разряда – в ячейку `H3` вводится
  `=E3-E2`
  нажатие Enter, маркер автозаполнения переводится от `H3` до `H12`;
- вычислить относительные частоты – в ячейку `I3` вводится
  `=F3/1000`
  нажатие Enter, маркер автозаполнения переводится от `I3` до `I12`; в данном примере число 1000 – это объем выборки;
- вычислить высоту каждого прямоугольника гистограммы – в ячейку `J3` вводится
  `=I3/H3`
  нажатие Enter, маркер автозаполнения переводится от ячейки `J3` до ячейки `J12`.

Далее из меню Вставка выбирается Диаграмма. На вкладке Стандартные выбирается Гистограмма. После перехода к следующему диалоговому окну (нажатие на Далее) на вкладке Диапазон данных в поле Диапазон указывается интервал ячеек, содержащий высоты прямоугольников (в данном примере – «`=Лист1!$J$3:$J$12`»). В этом же диалоговом окне на вкладке Ряд в поле Подписи оси X указывается интервал ячеек со значениями, соответствующими центру каждого разряда (в данном примере – «`=Лист1!$G$3:$G$12`»). В следующем диалоговом окне на вкладке Заголовки в поле Ось X (категорий) можно ввести строку «`Центр разряда`». В последнее диалоговое окно мастера диаграмм никакой информации вводить не нужно (выбирается **Далее**, затем – **Готово**); в результате будет построена гистограмма:



После этого можно изменить ширину каждого прямоугольника (двойной щелчок на любом из них, в диалоговом окне **Формат ряда данных** на вкладке **Параметры** установить значение в поле **Ширина зазора** равным 0 или 1) и удалить заголовок ряда (щелчок на заголовке «`Ряд 1`», затем – нажатие на **Delete**).

2. Основные законы распределения

В процессе решения статистических задач часто требуется выполнить сравнение двух величин, одна из которых вычисляется на основе выборочных характеристик (оценок среднего, дисперсии и т.д.), а другая является значением функции распределения одной из статистик (или квантилью этой статистики – значением функции, обратной к функции распределения).

Наиболее распространенные статистики являются моделями типичных задач теории вероятностей, возникающих в практических ситуациях.

В связи с задачей о совместном влиянии случайных величин возникает важнейшее распределение, называемое *нормальным*. Именно, если величина X является суммой большого числа независимых случайных величин, то плотность распределения величины X имеет вид

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} = N(m,\sigma), \qquad (2.1)$$

где m и σ – константы, равные математическому ожиданию и стандартному отклонению случайной величины X. Если $m = 0$ и $\sigma = 1$, то распределение называют *стандартным* (или *нормированным*) *нормальным распределением*.

График плотности вероятности (2.1) нормального распределения называется *нормальной кривой* (или *кривой Гаусса*). Выражение (2.1) определяет четную функцию относительно разности $x - m$, поэтому нормальная кривая симметрична относительно прямой $x = m$. Медиана и мода нормального распределения совпадают с математическим ожиданием. По мере удаления от точки $x = m$ плотность быстро уменьшается и при $x \to \pm\infty$ асимптотически приближается к нулю. При изменении математического ожидания m нормальная кривая смещается вдоль оси абсцисс, не изменяя своей формы. При уменьшении σ кривая становится более «островершинной», сжимаясь вдоль оси абсцисс; при увеличении σ кривая становится более «пологой».

Вероятность попадания нормально распределенной случайной величины на данный интервал

$$P(\alpha \le X < \beta) = \Phi\left(\frac{\beta - m}{\sigma}\right) - \Phi\left(\frac{\alpha - m}{\sigma}\right), \qquad (2.2)$$

где $\Phi(t)$ – *фЗнкцияяЛапласа*:я

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{u^2}{2}} du. \qquad (2.3)$$

Иногда функцией Лапласа называют функцию

$$2\Phi(t) = \sqrt{\frac{2}{\pi}} \int_0^t e^{-\frac{u^2}{2}} du.$$

Если (из таблиц) известно значение именно этой функции, то правую часть соотношения (2.2) необходимо разделить на 2.

Известным может оказаться значение *интегралаяошибок*:

$$erf(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-u^2} du.$$

Функция Лапласа связана с ним соотношением

$$\Phi(t) = \frac{1}{2} erf\left(\frac{t}{\sqrt{2}}\right). \qquad (2.4)$$

Начиная с tя= 2 можно применять асимптотическую формулу

$$\Phi(t) \approx \frac{1}{2} - \frac{1}{t\sqrt{2\pi}} e^{-\frac{t^2}{2}}. \qquad (2.5)$$

При $t = 2$ соотношение (2.5) дает абсолютную погрешность около 0,004; при $t = 3$ погрешность уменьшается до 10^{-4}.

Наиболее важную роль в математической статистике играет *распределение Пирсона*, иначе называемое χ^2*-распределением*. Этому распределению подчинена сумма квадратов k независимых случайных величин:

$$X = \sum_{i=1}^{k} Y_i^2, \qquad (2.6)$$

каждая из которых, в свою очередь, распределена по стандартному нормальному закону. Плотность χ^2-распределения

$$f(x) = \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} e^{-\frac{x}{2}} x^{\frac{k}{2}-1}, \qquad (2.7)$$

где $\Gamma(z)$ – *гамма-функци* :

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt. \qquad (2.8)$$

Графики плотности χ^2-распределения приведены на рис. 2.1.

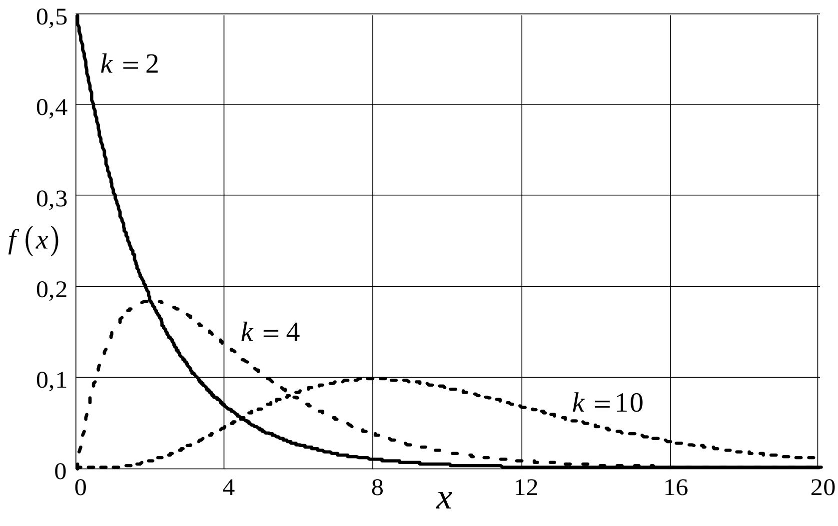


Рис. 2.1. Плотность χ^2-распределения для различного числа степеней свободы

С увеличением числа степеней свободы плотность (2.7) приближается к плотности нормального закона. Справедлива асимптотическая формула

$$F(x) = P(x < X) \to \Phi^*\left(\sqrt{2x}\right) - \Phi^*\left(\sqrt{2k-1}\right), \tag{2.9}$$

где $\Phi^*(t)$ – функция стандартного нормального распределения.

Распределением Стьюдента с k степенями свободы называется распределение случайной величины:

$$X = U\sqrt{\frac{1}{kY}}, \tag{2.10}$$

где U – случайная величина, подчиненная стандартному нормальному закону, Y – случайная величина, подчиненная χ^2-распределению с k степенями свободы.

Плотность распределения Стьюдента

$$f(x) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}. \tag{2.11}$$

Графики функции (2.11) для различного числа степеней свободы изображены на рис. 2.2.

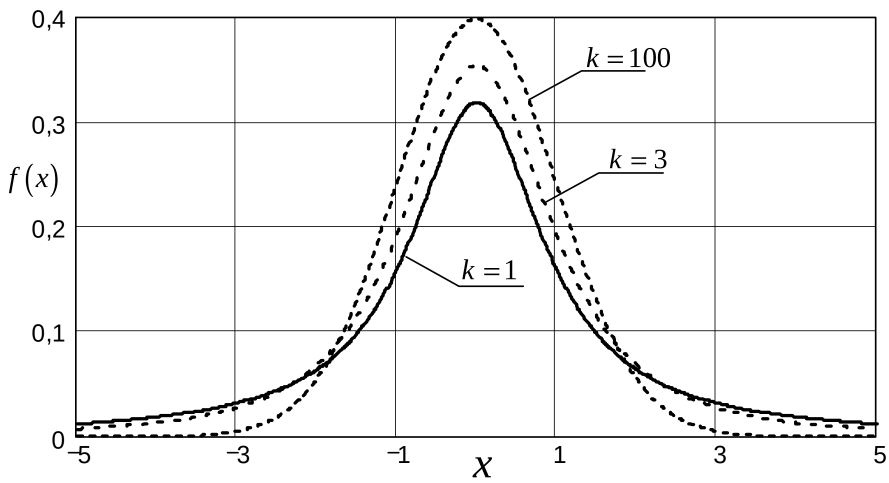


Рис. 2.2. Плотность распределения Стьюдента

Распределением Фишера, или *F-распределением* с m и n степенями свободы называется распределение случайной величины

$$X = \frac{n}{m}\frac{Y_m}{Y_n}, \tag{2.12}$$

где Y_m, Y_n – случайные величины, подчиненные χ^2-распределениям со степенями свободы m и n, соответственно.

Плотность F-распределения:

$$f(x)=\frac{1}{\mathrm{B}\left(\frac{m}{2},\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}}x^{\frac{m}{2}-1}\left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}}, \qquad (2.13)$$

где

$$\mathrm{B}(u,v)=\int_0^1 t^{u-1}(1-t)^{v-1}dt=\frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}$$

– *бета-функция.*

Графики (2.13) изображены на рис. 2.3.

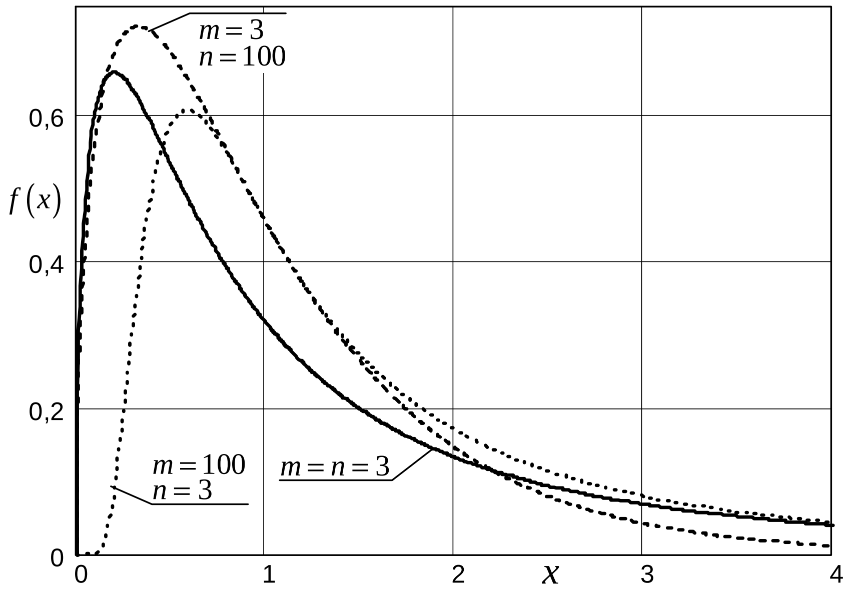


Рис. 2.3. Плотность F-распределения

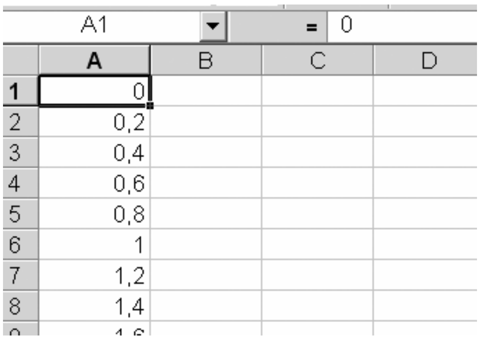
2.1. *Табулирование функции χ^2-распределения*

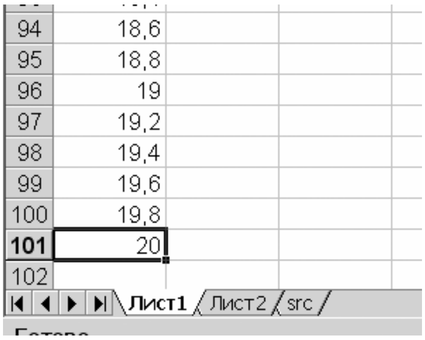
Использование статистических пакетов для табулирования функций распределения (и значений соответствующих квантилей) избавляет от необходимости обращения к таблицам.

Среди функций рабочего листа пакета `MS Excel` имеется функция `ХИ2ОБР`, возвращающая вероятность того, что подчиненная χ^2-распределению случайная величина X примет значение, *большее или равное* заданного x (по неизвестным причинам разработчики пакета проигнорировали общепринятое определение функции распределения как вероятности события $X < x$; это следует принять как данность).

Для построения графика функции χ^2-распределения (число степеней свободы выбрано равным $k = 4$) можно выполнить следующие действия:

1. Определить в первом столбце рабочего листа сетку. В простейшем случае сетка может быть равномерной, и для ее введения достаточно средств автозаполнения. Левой границей является 0. Положение правой границы зависит от числа степеней свободы; для $k \sim 4$ правой границей может быть $x_{\max} = 20$. Шаг сетки можно выбрать равным $0{,}1 \le \Delta x \le 0{,}4$.

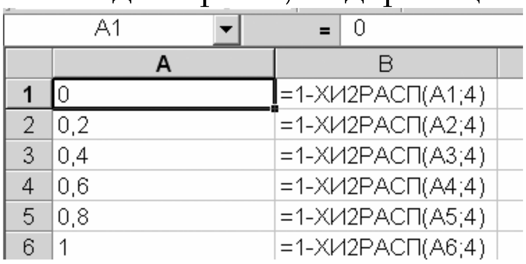


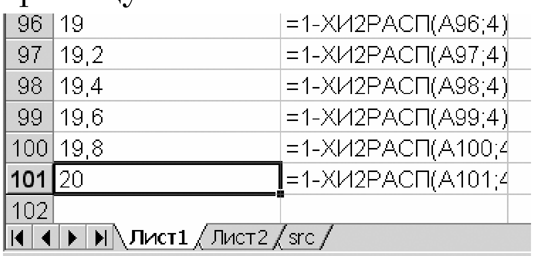


2. В первую строку второго столбца ввести формулу

```
=1-ХИ2РАСП(A1;4)
```

здесь `A1` – ссылка на первую ячейку столбца, содержащего значения случайной величины, 4 – выбранное в данном примере число степеней свободы. После этого следует перевести маркер автозаполнения до строки, содержащей правую границу сетки.

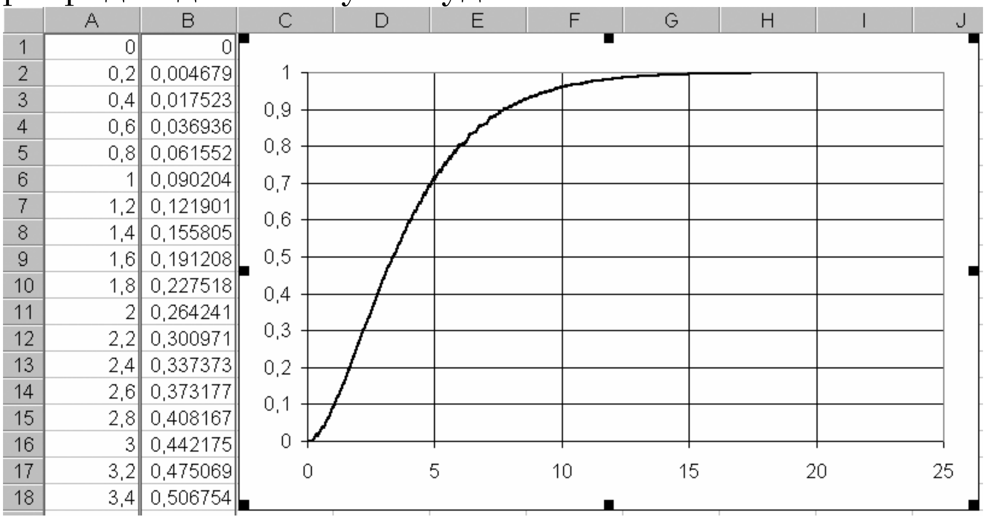




3. В двух первых столбцах выделить диапазон строк, содержащий сетку и табулированные значения функции распределения. Из меню **Вставка** выбрать **Диаграмма**, затем – двойной щелчок на пункте **Точечная**. В следующем диалоговом окне на вкладках **Диапазон данных** (в поле **Диапазон**) и **Ряд** уже должны находиться корректные значения; менять их не следует. На шаге 3 можно добавить назва-

ния осей и установить флажки линий сетки. На шаге 4 в качестве назначения можно выбрать текущий рабочий лист.

4. Нужно проследить корректность пределов оси ординат: минимальное значение должно быть равно 0, максимальное – равно 1. Маркеры рядов данных лучше удалить.



	A	B
1	0	0
2	0,2	0,004679
3	0,4	0,017523
4	0,6	0,036936
5	0,8	0,061552
6	1	0,090204
7	1,2	0,121901
8	1,4	0,155805
9	1,6	0,191208
10	1,8	0,227518
11	2	0,264241
12	2,2	0,300971
13	2,4	0,337373
14	2,6	0,373177
15	2,8	0,408167
16	3	0,442175
17	3,2	0,475069
18	3,4	0,506754

2.2. Табулирование функции распределения Стьюдента

Если в (2.10) число степеней свободы k является целым, то и плотность вероятности (2.11), и функция распределения Стьюдента являются элементарными. Однако даже в этом случае выражения указанных функций весьма громоздки и вычисление их значений затруднительно.

Среди функций рабочего листа пакета `MS Excel` имеется функция `СТЬЮДРАСП`, принимающая три параметра: значение случайной величины, для которого отыскивается соответствующее значение функции распределения; число степеней свободы и число «хвостов». Целесообразность введения последнего параметра связана с тем, что распределение Стьюдента симметрично и часто используется в *двусторонних* оценках. Имеет место равенство

```
2*СТЬЮДРАСП(x,k,1) = СТЬЮДРАСП(x,k,2)
```

При вызове

```
=СТЬЮДРАСП(x,k,1)
```

возвращается число

$$P(X \ge x) = 1 - \int_{-\infty}^{x} f(t,k)\,dt$$

– вероятность того, что подчиненная распределению Стьюдента с k степенями свободы случайная величина X примет значение, *большее* *или равное* заданного x. Кроме этого, на значение x (по причине, известной лишь разработчикам Excel) накладывается ограничение $x \geq 0$.

Для построения графика функции распределения Стьюдента (число степеней свободы $k = 4$) можно выполнить следующие действия:

1. Определить в первом столбце рабочего листа сетку; в данном примере – равномерная сетка от −5 до 5 с шагом 0,2.

A1 = -5

	A	B	C
1	-5		
2	-4,8		
3	-4,6		
4	-4,4		
5	-4,2		
6	-4		
7	-3,8		
8	-3,6		

A51 = 5

	A	B	C
44	3,6		
45	3,8		
46	4		
47	4,2		
48	4,4		
49	4,6		
50	4,8		
51	5		

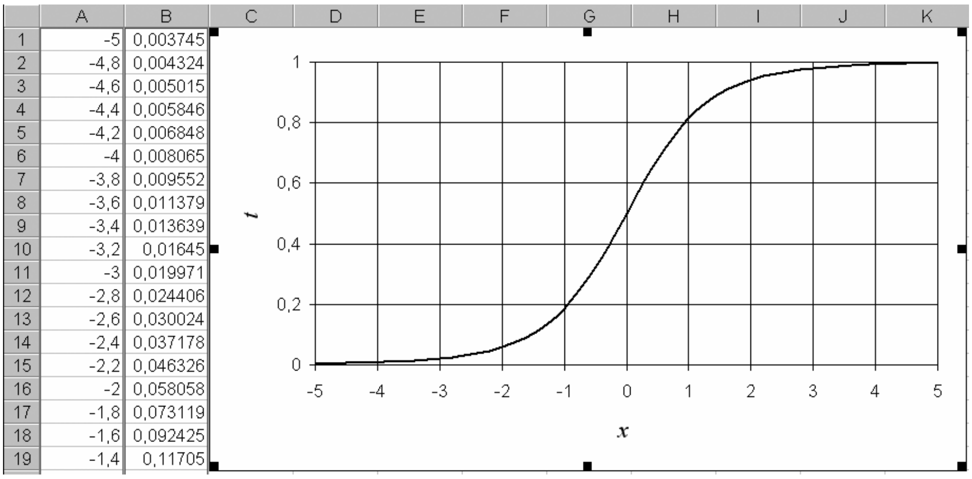
2. В первую строку второго столбца ввести формулу
`=ЕСЛИ(A1<0;СТЬЮДРАСП(-A1;4;1);1-СТЬЮДРАСП(A1;4;1))`

Здесь функция ЕСЛИ использована для обхода ограничения $x \geq 0$.

	A	B	C	D	E	F	G
1	-5	=ЕСЛИ(A1<0;СТЬЮДРАСП(-A1;4;1);1-СТЬЮДРАСП(A1;4;1))					
2	-4,8						
3	-4,6						
4	-4,4						

После этого следует перевести маркер автозаполнения до строки, содержащей правую границу сетки.

3. Как и в предыдущем примере, в двух первых столбцах выделить диапазон строк, содержащий сетку и табулированные значения функции распределения. Из меню **Вставка** выбрать **Диаграмма**, затем – двойной щелчок на пункте **Точечная**. На вкладках **Диапазон данных** и **Ряд** уже должны находиться корректные значения. На шаге 3 можно добавить названия осей и установить флажки линий сетки. На шаге 4 в качестве назначения можно выбрать текущий рабочий лист. Следует проследить корректность пределов осей.



	A	B
1	-5	0,003745
2	-4,8	0,004324
3	-4,6	0,005015
4	-4,4	0,005846
5	-4,2	0,006848
6	-4	0,008065
7	-3,8	0,009552
8	-3,6	0,011379
9	-3,4	0,013639
10	-3,2	0,01645
11	-3	0,019971
12	-2,8	0,024406
13	-2,6	0,030024
14	-2,4	0,037178
15	-2,2	0,046326
16	-2	0,058058
17	-1,8	0,073119
18	-1,6	0,092425
19	-1,4	0,11705

2.3. Генерация случайных чисел, подчиненных данному закону

При выполнении имитационного моделирования могут потребоваться выборки из генеральных совокупностей, подчиненных заранее заданным законам распределения. Стандартные библиотеки большинства алгоритмических языков (как, впрочем, и функция СЛЧИС рабочего листа `MS Excel`) позволяют получать выборки, подчиненные лишь закону равномерной плотности.

Используя средства пакета анализа `MS Excel`, можно получить выборки, подчиненные биномиальному, нормальному распределениям и распределению Пуассона[1].

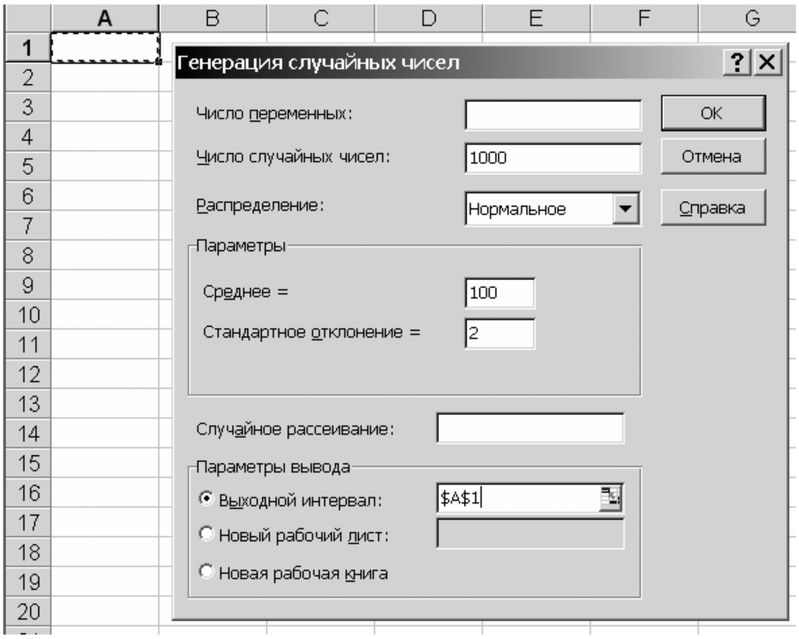
Пусть требуется получить 1000 вариант, подчиненных нормальному распределению с параметрами $M[X]=100$, $D[X]=4$ (стандартное отклонение $\sigma=\sqrt{D[X]}=2$). Можно выполнить следующие действия.

1. Из меню **Сервис** выбрать **Анализ данных**, далее – **Генерация случайных чисел**. В поле **Число переменных** следует указать число столбцов, а в поле **Число случайных чисел** – число строк, которые нужно заполнить вариантами. Пусть все варианты требуется поместить в первый столбец (заполнить 1000 строк); тогда первое поле можно оставить пустым, а во второе следует ввести значение `1000`.

2. В списке **Распределение** выбрать **Нормальное**; установить требуемые параметры (выберем **Среднее** равным `100`, **Стандартное отклонение** – равным `2`). В поле **Выходной интервал** указать `$A$1` – первую ячейку первого столбца. Подтвердить ввод.

---

[1] Следует помнить, что генерируемые цифровой ЭВМ числа в действительности *неслучайны*; однако используемые для их генерации алгоритмы таковы, что полученные значения практически могут считаться случайными.



Варианты будут помещены в ячейки `A1` – `A1000`.

Для получения 1000 вариант, подчиненных χ^2-распределению с четырьмя степенями свободы, можно выполнить следующие действия.

1. Заполнить первые четыре столбца случайными значениями, подчиненными стандартному закону (аналогично предыдущему примеру, но в поле **Число переменных** нужно указать `4`; в полях **Среднее** и **Стандартное отклонение** указываются значения `0` и `1`, соответственно).

2. Заполнить пятый столбец суммой квадратов значений, расположенных в соответствующих строках первых четырех столбцов.

Последнее действие можно выполнить с использованием автозаполнения. В ячейку `E1` вводится формула:

```
=A1^2+B1^2+C1^2+D1^2
```

После этого маркер автозаполнения следует переместить до строки с номером 1000.

3. Проверка статистических гипотез

Одной из основных задач математической статистики является проверка некоторых суждений на основе опытных данных. Например, в прикладных задачах очень часто ставится вопрос о наличии так называемого *эффекта обработки*; этот вопрос может быть сформулирован по-разному.

- верно ли, что смена технологии позволяет получить продукцию лучшего качества (предполагается, что сформулирован какой-либо критерий качества);
- верно ли, что смена технологии приводит к уменьшению «разброса» значений некоторого показателя (уменьшению его коэффициента вариации).

Исходные данные, на основе которых решаются подобные вопросы, обычно бывают получены опытным путем, в результате выборочного обследования продукции (замеров ее показателей). Поэтому ответ на вопрос может быть дан *лишь с определенной степенью уверенности*; существует некоторая ненулевая вероятность ошибки. Задача математической статистики – выработать методы, которые позволяют оценить вероятность этой ошибки.

Пусть в результате выборочного обследования получено наблюдение (извлечена выборка) $\mathbf{x}$. Пусть X – множество всевозможных наблюдений (*выборочное пространство*; это понятие не следует смешивать с понятием генеральной совокупности). Появление наблюдения $\mathbf{x}$ происходит в соответствии с некоторым распределением вероятности на выборочном пространстве (некоторые наблюдения более вероятны, нежели другие).

Статистической гипотезой H называется предположение о виде неизвестного распределения или о параметрах известного распределения вероятностей на выборочном пространстве.

Конкурирующей (*альтернативной*) гипотезой H_1 называют гипотезу, противоположную гипотезе H_0; по отношению к конкурирующей гипотезу H_0 называют *исходной* (*нулевой*).

Большинство статистических гипотез может быть сформулировано в одной из следующих двух форм.

1. Данные выборки получены из генеральных совокупностей с равными математическими ожиданиями (или равными моментами высших порядков).
2. Данная выборка извлечена из генеральной совокупности, подчиненной определенному распределению.

Проверка статистической гипотезы H состоит в выяснении того, насколько эта гипотеза согласуется с опытными данными $\mathbf{x}$. Содержание этой операции сводится к *неформальному* выбору связанного с наблюдением $\mathbf{x}$ события A, и последующему *формальному* отысканию его вероятности $P(A|H)$ при гипотезе H. Событие A принято выбирать так, чтобы его вероятность $P(A|H)$ оказалась малой; в этом случае A называют *критическим событием*, или *статистическим критерием* для гипотезы H. Более строго – *статистическим критерием*, или *статистикой*, называют случайную величину (с известным распределением), которая служит для проверки гипотезы.

Если вероятность $P(A|H)$ *реально наблюдаемого* в опыте критического события A оказывается меньше некоторого заранее заданного *уровняязначимости* α, то *гипотезаяНяотвергаетсяянаяуровнея значимостия*α. Предположим, что гипотеза H верна (т.е. является достоверным событием): $P(H|A) = P(H) = 1$. По определению условной вероятности

$$P(A|H) = \frac{P(A)P(H|A)}{P(H)} = P(A);$$

таким образом, уровень значимости – это вероятность того, что *верная я гипотезая будетя ошибочноя отвергнута*. Соответствующую ошибку, состоящую в непринятии правильной на самом деле гипотезы, называют *ошибкойяпервогоярода*.

Важно, что методы математической статистики *неяпозволяют* получить ответ на вопрос об истинности гипотезы. Они лишь дают возможность сделать *вероятностное* суждение, позволяющее ее опровергнуть.

Ошибкойявторогоярода называют ошибку, состоящую в том, что принимается неверная гипотеза (альтернатива для верной). Вероятность ошибки второго рода дополняет до единицы число, называемое *мощностьюя*статистического критерия. Обычно одна и та же гипотеза может быть проверена при помощи различных критериев. Среди этих критериев следует по возможности выбирать тот, который обладает наибольшей мощностью: *значенияякритерияяприянулевойягипотезеяияальтернативеядолжныяотличатьсяякакяможноя больше*.

Выбор уровня значимости отражает принятую малую вероятность события, которое на практике считается *невозможным*. В большинстве поисковых исследований уровень значимости α выбирают равным 0,05. Критические задачи (например, связанные с оценкой надежности транспортных средств) требуют выбора существенно меньших значений: $\alpha \sim 10^{-6}$; однако при столь малом уровне значимости большинство методов математической статистики становятся непригодными для использования.

В простейших задачах критическое событие A можно отождествить с *непо влением наблюдени* **x**. Если для подобной формулировки удается отыскать вероятность $P(A|H)$, то говорят, что гипотеза проверяется *непосредственно*.

На практике вероятность появления (или непоявления) данного наблюдения **x** при гипотезе H отыскать обычно не удается. Поэтому ограничиваются проверкой следствий, вытекающих из содержания гипотезы.

4. Гипотеза о нормальном распределении генеральной совокупности

Одной из наиболее распространенных *одновыборочных* задач является проверка гипотезы о нормальном распределении генеральной совокупности.

При использовании χ^2-*статистики* после построения непрерывного вариационного ряда вычисляется значение статистики (случайной величины, связанной с опытными данными):

$$\chi^2 = \frac{1}{n}\sum_{j=1}^{l}\frac{\left(n_j - np_j\right)^2}{p_j}, \qquad (4.1)$$

где n – объем выборки (как правило – не менее 200); l – число разрядов непрерывного вариационного ряда (не менее 8); n_j – частота; p_j – вероятность, найденная расчетом по нормальной кривой, выравнивающей выборку:

$$p_j = \frac{1}{s\sqrt{2\pi}}\int_{x_j}^{x_{j+1}} e^{-\frac{(x-\bar{x})^2}{2s^2}}dx = \Phi\left(\frac{x_{j+1}-\bar{x}}{s}\right) - \Phi\left(\frac{x_j-\bar{x}}{s}\right), \;\; j=\overline{1,l}. \qquad (4.2)$$

В последнем соотношении $\bar{x}$ – оценка математического ожидания, s – оценка среднего квадратичного отклонения, $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$ – функция Лапласа (или функция стандартного нормального распределения).

Статистика (4.1) подчинена χ^2-распределению с числом степеней свободы $l-3$. С учетом этого ищется вероятность критического события, состоящего в том, что для выборки из нормально распределенной генеральной совокупности истинное (неизвестное) значение статистики окажется столь же большим (большим или равным), как и наблюдаемое на опыте значение. Если указанная вероятность близка к нулю (меньше выбранного уровня значимости α), то нулевая гипотеза о нормальном распределении генеральной совокупности отвергается.

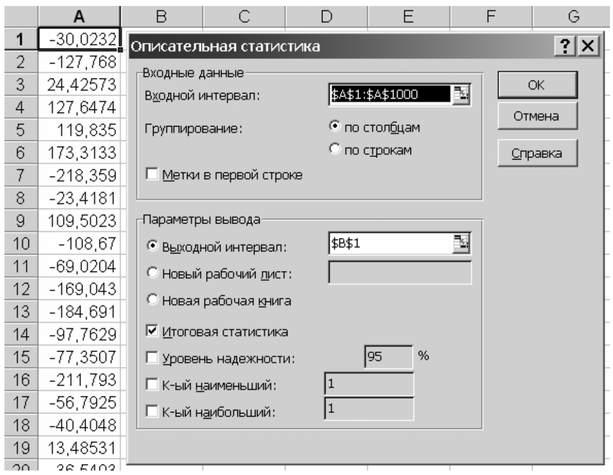
Вместо нахождения вероятности критического события можно сравнить найденное значение статистики (4.1) с квантилью распределения $\chi^2_{l-3,\alpha}$ для $l-3$ степеней свободы и выбранного уровня значимости α. При выполнении неравенства

$$\chi^2 < \chi^2_{l-3,\alpha}$$

считают, что на уровне значимости α нет оснований отвергать нулевую гипотезу о нормальном распределении генеральной совокупности (вновь отметим, что *истинность* гипотезы этим не доказывается!).

4.1. Использование средств MS Excel для проверки гипотезы о нормальном распределении генеральной совокупности

Пусть требуется проверить гипотезу о нормальном распределении генеральной совокупности, из которой извлечена выборка с объемом 1000. Пусть варианты помещены в первый столбец рабочего листа (ячейки `A1:A1000`).



Для нахождения оценок среднего и стандартного отклонения, а также для построения непрерывного вариационного ряда можно воспользоваться пакетом анализа.

Из меню Сервис выбрать Анализ данных, далее – Описательная статистика. Входной интервал – `$A$1:$A$1000`; верхняя левая ячейка выходного интервала `$B$1`. Установить флаг Итоговая статистика.

Для данной выборки минимальное и максимальное значения оказались равными −277 и 302 соответственно, поэтому в качестве границ первого и последнего разрядов можно выбрать −280 и 310:

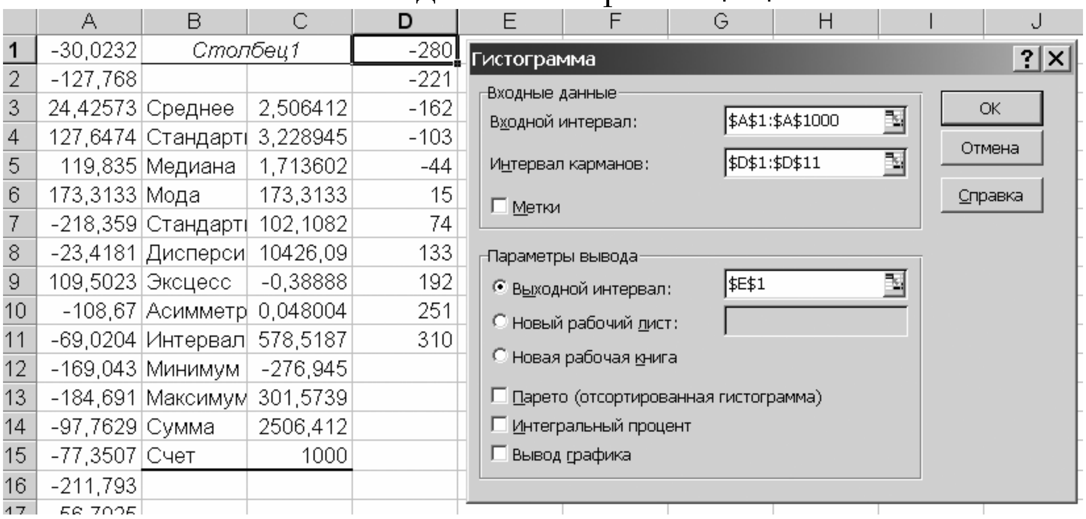
12	-169,043	Минимум	-276,945
13	-184,691	Максимум	301,5739

Так как $1+3{,}3\lg 1000=12{,}9$, то число разрядов можно взять равным 10. Тогда длина разряда

$$\frac{310-(-280)}{10}=59\,.$$

В столбце `D` введем равномерную сетку от `–280` до `310` с шагом `59` (можно воспользоваться средствами автозаполнения: ввести в ячейку `D1` значение `–280`, в ячейку `D2` – формулу `=D1+59`, выделить ячейки `D1:D2` и переместить маркер автозаполнения до ячейки `D11`).

Для построения непрерывного вариационного ряда из меню **Сервис** следует выбрать **Анализ данных**, далее – **Гистограмма**. Входной интервал – `$A$1:$A$1000`; границы разрядов – `$D$1:$D$11`; верхняя левая ячейка выходного интервала `$E$1`.



	A	B	C	D
1	-30,0232	*Столбец1*		-280
2	-127,768			-221
3	24,42573	Среднее	2,506412	-162
4	127,6474	Стандарті	3,228945	-103
5	119,835	Медиана	1,713602	-44
6	173,3133	Мода	173,3133	15
7	-218,359	Стандарті	102,1082	74
8	-23,4181	Дисперси	10426,09	133
9	109,5023	Эксцесс	-0,38888	192
10	-108,67	Асимметр	0,048004	251
11	-69,0204	Интервал	578,5187	310
12	-169,043	Минимум	-276,945	
13	-184,691	Максимум	301,5739	
14	-97,7629	Сумма	2506,412	
15	-77,3507	Счет	1000	
16	-211,793			

Найдем вероятности p_j, $j=\overline{1{,}10}$, соответствующие нормальной кривой, выравнивающей выборку. Для каждой из одиннадцати границ x_j, $j=\overline{1{,}11}$ можно вычислить значение

$$x_j^*=\frac{x_j-\bar{x}}{s},$$

соответствующее «стандартному» положению разрядов. В ячейку `G2` следует ввести формулу

```
=(E2-$C$3)/$C$7
```

и переместить маркер автозаполнения до ячейки `G12` (обратить внимание – запись ссылок `$C$3` и `$C$7` на ячейки, содержащие среднее и оценку стандартного отклонения, говорит о том, что эти ссылки *не следует изменять в процессе автозаполнения*).

	A	B	C	D	E	F	G	H
1	-30,0232	*Столбец1*		-280	*Карман*	*Частота*		
2	-127,768			-221	-280	0	=(E2-$C$3)/$C$7	
3	24,42573	Среднее	2,506412	-162	-221	6	-2,18892	
4	127,6474	Стандартн	3,228945	-103	-162	51	-1,6111	
5	119,835	Медиана	1,713602	-44	-103	94	-1,03328	
6	173,3133	Мода	173,3133	15	-44	186	-0,45546	
7	-218,359	Стандартн	102,1082	74	15	217	0,122356	
8	-23,4181	Дисперси	10426,09	133	74	191	0,700175	
9	109,5023	Эксцесс	-0,38888	192	133	151	1,277993	
10	-108,67	Асимметр	0,048004	251	192	75	1,855811	
11	-69,0204	Интервал	578,5187	310	251	24	2,43363	
12	-169,043	Минимум	-276,945		310	5	3,011448	
13	-184,691	Максимум	301,5739		Еще	0		
14	-97,7629	Сумма	2506,412					
15	-77,3507	Счет	1000					

Для нахождения вероятностей p_j поместим в ячейку H3 формулу

=НОРМСТРАСП(G3)-НОРМСТРАСП(G2)

и переместим маркер автозаполнения до ячейки H12.

	A	B	C	D	E	F	G	H	I	J	K
1	-30,0232	*Столбец1*		-280	*Карман*	*Частота*					
2	-127,768			-221	-280	0	-2,76674				
3	24,42573	Среднее	2,506412	-162	-221	6	-2,18892	=НОРМСТРАСП(G3)-НОРМСТРАСП(G2)			
4	127,6474	Стандартн	3,228945	-103	-162	51	-1,6111	0,039278			
5	119,835	Медиана	1,713602	-44	-103	94	-1,03328	0,097157			
6	173,3133	Мода	173,3133	15	-44	186	-0,45546	0,173652			
7	-218,359	Стандартн	102,1082	74	15	217	0,122356	0,224303			
8	-23,4181	Дисперси	10426,09	133	74	191	0,700175	0,209399			
9	109,5023	Эксцесс	-0,38888	192	133	151	1,277993	0,141283			
10	-108,67	Асимметр	0,048004	251	192	75	1,855811	0,068886			
11	-69,0204	Интервал	578,5187	310	251	24	2,43363	0,024266			
12	-169,043	Минимум	-276,945		310	5	3,011448	0,006174			
13	-184,691	Максимум	301,5739		Еще	0					
14	-97,7629	Сумма	2506,412								
15	-77,3507	Счет	1000								

Вычислим сумму

$$\chi^2 = \frac{1}{n}\sum_{j=1}^{l}\frac{\left(n_j - np_j\right)^2}{p_j}.$$

В ячейку I3 введем

=(F3-$C$15*H3)^2/H3

и переместим маркер автозаполнения до ячейки I12.

	A	B	C	D	E	F	G	H	I	J
1	-30,0232	*Столбец1*		-280	*Карман*	*Частота*				
2	-127,768			-221	-280	0	-2,76674			
3	24,42573	Среднее	2,506412	-162	-221	6	-2,18892	0,01147	=(F3-$C$15*H3)^2/H3	
4	127,6474	Стандартн	3,228945	-103	-162	51	-1,6111	0,039278	3498,474	
5	119,835	Медиана	1,713602	-44	-103	94	-1,03328	0,097157	102,6005	
6	173,3133	Мода	173,3133	15	-44	186	-0,45546	0,173652	878,0246	
7	-218,359	Стандартн	102,1082	74	15	217	0,122356	0,224303	237,7873	
8	-23,4181	Дисперси	10426,09	133	74	191	0,700175	0,209399	1616,699	
9	109,5023	Эксцесс	-0,38888	192	133	151	1,277993	0,141283	668,3008	
10	-108,67	Асимметр	0,048004	251	192	75	1,855811	0,068886	542,6798	
11	-69,0204	Интервал	578,5187	310	251	24	2,43363	0,024266	2,916236	
12	-169,043	Минимум	-276,945		310	5	3,011448	0,006174	223,2538	
13	-184,691	Максимум	301,5739		Еще	0				

После этого в ячейку I13 введем

=СУММ(I3:I12)/$C$15

	A	B	C	D	E	F	G	H	I	J
1	-30,0232	*Столбец1*		-280	*Карман*	*Частота*				
2	-127,768			-221	-280	0	-2,76674			
3	24,42573	Среднее	2,506412	-162	-221	6	-2,18892	0,01147	2608,833	
4	127,6474	Стандарт	3,228945	-103	-162	51	-1,6111	0,039278	3498,474	
5	119,835	Медиана	1,713602	-44	-103	94	-1,03328	0,097157	102,6005	
6	173,3133	Мода	173,3133	15	-44	186	-0,45546	0,173652	878,0246	
7	-218,359	Стандарт	102,1082	74	15	217	0,122356	0,224303	237,7873	
8	-23,4181	Дисперси	10426,09	133	74	191	0,700175	0,209399	1616,699	
9	109,5023	Эксцесс	-0,38888	192	133	151	1,277993	0,141283	668,3008	
10	-108,67	Асимметр	0,048004	251	192	75	1,855811	0,068886	542,6798	
11	-69,0204	Интервал	578,5187	310	251	24	2,43363	0,024266	2,916236	
12	-169,043	Минимум	-276,945		310	5	3,011448	0,006174	223,2538	
13	-184,691	Максимум	301,5739		Еще	0			=СУММ(I3:I12)/$C$15	

Число разрядов $l = 10$, поэтому число степеней свободы χ^2-распределения равно $10 - 3 = 7$. Для вычисления вероятности критического события, состоящего в том, что значение случайной величины, подчиненной χ^2-распределению, окажется столь же большим, как и наблюдаемое на опыте значение, в ячейку `I14` введем

`=ХИ2РАСП(I13;7)`

	A	B	C	D	E	F	G	H	I
13	-184,691	Максимум	301,5739		Еще	0			10,37957
14	-97,7629	Сумма	2506,412		Вероятность критического события:				=ХИ2РАСП(I13;7)

Окончательный результат:

	A	B	C	D	E	F	G	H	I
1	-30,0232	*Столбец1*		-280	*Карман*	*Частота*			
2	-127,768			-221	-280	0	-2,76674		
3	24,42573	Среднее	2,506412	-162	-221	6	-2,18892	0,01147	2608,833
4	127,6474	Стандарт	3,228945	-103	-162	51	-1,6111	0,039278	3498,474
5	119,835	Медиана	1,713602	-44	-103	94	-1,03328	0,097157	102,6005
6	173,3133	Мода	173,3133	15	-44	186	-0,45546	0,173652	878,0246
7	-218,359	Стандарт	102,1082	74	15	217	0,122356	0,224303	237,7873
8	-23,4181	Дисперси	10426,09	133	74	191	0,700175	0,209399	1616,699
9	109,5023	Эксцесс	-0,38888	192	133	151	1,277993	0,141283	668,3008
10	-108,67	Асимметр	0,048004	251	192	75	1,855811	0,068886	542,6798
11	-69,0204	Интервал	578,5187	310	251	24	2,43363	0,024266	2,916236
12	-169,043	Минимум	-276,945		310	5	3,011448	0,006174	223,2538
13	-184,691	Максимум	301,5739		Еще	0			10,37957
14	-97,7629	Сумма	2506,412		Вероятность критического события:				0,168064
15	-77,3507	Счет	1000						

Пусть уровень значимости выбран равным $\alpha = 0{,}05$. Так как $0{,}168 > 0{,}05$, то на данном уровне значимости гипотеза о нормальном распределении генеральной совокупности не противоречит опытным данным.

5. Некоторые двухвыборочные задачи

На практике часто встречается случай, когда средний результат одной серии экспериментов отличается от среднего результата другой серии. При этом возникает вопрос, является ли обнаруженное расхождение средних *статистически значимым* – можно ли объяснить его случайными ошибками или же оно имеет закономерное значение. В промышленности задача сравнения средних часто возникает при контроле качества продукции, изготовленной при различных технологических режимах.

Задача сравнения средних решается различно в зависимости от того, являются ли известными дисперсии двух совокупностей (и если они известны – то в зависимости от того, равны ли они). Очевидно, что значимость различия средних зависит от дисперсий генеральных совокупностей – малость различия в сравнении со стандартным отклонением указывает на его незначимость. Однако тогда, когда генеральные средние оцениваются по результатам эксперимента (т.е. заменяются выборочными средними), то различие между средними может быть значимым даже в том случае, если оно мало по сравнению со стандартным отклонением (указанная ситуация имеет место для выборок большого объема). Именно по этой причине «количественной характеристикой» различия между средними является *стандартнаяяошибка* – частное от деления стандартного отклонения на корень квадратный из объема выборки (следует вспомнить, что дисперсия среднего из n независимых слагаемых в n раз меньше дисперсии каждого из них).

5.1. Проверка гипотезы о равенстве средних: случай известных и равных дисперсий

Наиболее просто задача сравнения генеральных средних $M\left[X\right]=\bar{x}_0$ и $M\left[Y\right]=\bar{y}_0$ решается в том случае, если дисперсии генеральных совокупностей, из которых извлечены выборки

$$\left\{x_i\right\},\ \left\{y_j\right\},\ i=\overline{1,N_1},\ j=\overline{1,N_2}$$

известны и равны σ_x^2 и σ_y^2, соответственно. Тогда можно принять, что выборочные средние $\bar{x}$ и $\bar{y}$ подчинены нормальным распределениям $N\left(\bar{x}_0,\sigma_x\right)$ и $N\left(\bar{y}_0,\sigma_y\right)$ – распределениям с плотностью

$$f\left(x\right)=\frac{1}{\sigma_x\sqrt{2\pi}}\exp\left(-\frac{\left(x-\bar{x}_0\right)^2}{2\sigma_x^2}\right),\ f\left(y\right)=\frac{1}{\sigma_y\sqrt{2\pi}}\exp\left(-\frac{\left(y-\bar{y}_0\right)^2}{2\sigma_y^2}\right).$$

Пусть проверяется гипотеза H_0 о равенстве генеральных средних. В случае справедливости этой гипотезы случайная величина, равная разности $\bar{x}-\bar{y}$ выборочных средних, подчинена нормальному закону распределения с математическим ожиданием

$$M[\bar{x}-\bar{y}]=M[\bar{x}]-M[\bar{y}]=0$$

и дисперсией

$$D[\bar{x}-\bar{y}]=D[\bar{x}]+D[\bar{y}]=\frac{\sigma_x^2}{N_1}+\frac{\sigma_y^2}{N_2}.$$

В последнем соотношении слагаемые в правой части представляют собой ни что иное, как *квадраты соответствующих стандартных ошибок*.

Так как неслучайный множитель можно выносить за знак дисперсии, возводя его в квадрат:

$$D[\alpha X]=\alpha^2 D[X],$$

то связанная с разностью $\bar{x}-\bar{y}$ выборочных средних статистика

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{D[\bar{x}-\bar{y}]}}=(\bar{x}-\bar{y})\left(\frac{\sigma_x^2}{N_1}+\frac{\sigma_y^2}{N_2}\right)^{-\frac{1}{2}}$$

подчинена стандартному нормальному закону $N(0,1)$.

Пусть в качестве конкурирующей гипотезы выбрана гипотеза H_1, состоящая в том, что для генеральных средних имеет место неравенство

$$M[X]\neq M[Y].$$

Тогда критическое событие A состоит в том, что случайная величина, подчиненная стандартному нормальному закону, окажется *не принадлежащей* интервалу $(-|t|;|t|)$. Вероятность этого события

$$P(A)=1-\frac{1}{\sqrt{2\pi}}\int\limits_{-|t|}^{-|t|}e^{-\frac{x^2}{2}}dx=1-2\Phi(|t|)=2-2\Phi^*(|t|),$$

где Φ – функция Лапласа, Φ^* – функция стандартного нормального распределения.

Если вероятность $P(A)$ оказывается меньше заранее заданного уровня значимости, то гипотеза H_0 о равенстве генеральных средних отвергается.

5.2. Проверка гипотезы о равенстве средних: случай неизвестных равных дисперсий

Пусть дисперсии генеральных совокупностей, из которых извлечены выборки $\{x_i\}$ и $\{y_j\}$ неизвестны (но предполагаются равными). Решение задачи сравнения генеральных средних начинается с вычисления *смешанной* оценки дисперсии разности выборочных средних:

$$D[\bar{x}-\bar{y}]=\frac{1}{N_1+N_2-2}\left(\frac{1}{N_1}+\frac{1}{N_2}\right)\left(\sum_{i=1}^{N_1}(x_i-\bar{x})^2+\sum_{i=1}^{N_2}(y_i-\bar{y})^2\right).$$

После этого находится эмпирическое значение статистики

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{D[\bar{x}-\bar{y}]}}. \qquad (5.1)$$

Статистика (5.1) подчинена распределению Стьюдента с $k=N_1+N_2-2$ степенями свободы. При альтернативе $M[X]\neq M[Y]$ вероятность критического события находится по формуле

$$P(A)=1-\frac{1}{\mathrm{B}\left(\frac{1}{2},\frac{k}{2}\right)\sqrt{k}}\int_{-|t|}^{|t|}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}dx,$$

где

$$\mathrm{B}(u,v)=\frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}=\int_0^1 x^{u-1}(1-x)^{v-1}dx$$

− бета-функция.

Зависимость вероятности критического события от значения модуля статистики t для выборок объемом $N_1=N_2=100$ приведена на рис. 5.1.

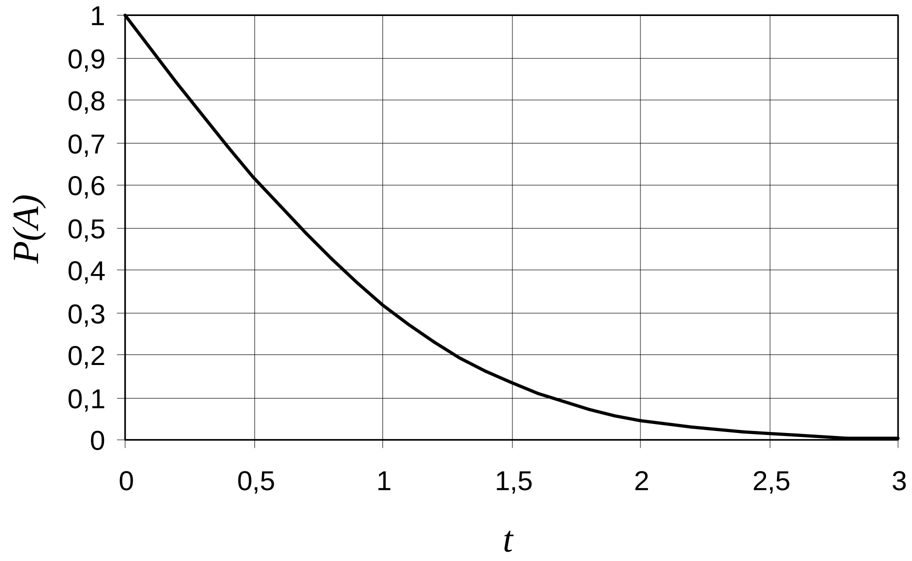


Рис. 5.1. Вероятность критического события
в задаче сравнения генеральных средних

Если вероятность $P(A)$ оказывается меньше заранее заданного уровня значимости, то гипотеза H_0 о равенстве генеральных средних отвергается в пользу альтернативы $M[X] \neq M[Y]$.

5.3. Проверка гипотезы о равенстве средних: случай неизвестных дисперсий

Если дисперсии генеральных совокупностей неизвестны и не предполагаются равными, то можно *приближенно* считать, что статистика

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{D[\bar{x} - \bar{y}]}},$$

где

$$D[\bar{x} - \bar{y}] = \frac{1}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\sum_{i=1}^{N_1}(x_i - \bar{x})^2 + \sum_{i=1}^{N_2}(y_i - \bar{y})^2\right),$$

также подчинена распределению Стьюдента. Однако соответствующее число степеней свободы уже не является целым числом и определяется достаточно сложным образом:

$$k = \frac{\left(\frac{s_x}{N_1} + \frac{s_y}{N_2}\right)^2}{\frac{s_x^2}{N_1^2(N_1 - 1)} + \frac{s_y^2}{N_2^2(N_2 - 1)}},$$

где

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{N_1 - 1}\sum_{i=1}^{N_1}(x_i - \bar{x})^2}\ ,\ \ s_y = \sqrt{s_y^2} = \sqrt{\frac{1}{N_2 - 1}\sum_{i=1}^{N_2}(y_i - \bar{y})^2}$$

– несмещенные оценки стандартных отклонений.

5.4. Проверка гипотезы о равенстве дисперсий

Дисперсия признака, как и связанные с ней характеристики – стандартное отклонение и коэффициент вариации – характеризуют такие исключительно важные показатели, как точность машин, приборов, технологических процессов и т.д.

Пусть имеются две нормально распределенные генеральные совокупности с неизвестными дисперсиями σ_x^2 и σ_y^2. Необходимо проверить нулевую гипотезу H_0: $\sigma_x^2 = \sigma_y^2$ о равенстве дисперсий.

Задача проверки сводится к сравнению несмещенных оценок генеральных дисперсий

$$s_x^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1}(x_i - \bar{x})^2\ ,\ \ s_y^2 = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2}(y_i - \bar{y})^2\ .$$

В случае справедливости нулевой гипотезы статистика, равная отношению этих оценок

$$F = \frac{s_x^2}{s_y^2}\ , \tag{5.2}$$

подчинена распределению Фишера (F-распределению) с числом степеней свободы $N_1 - 1$, $N_2 - 1$.

Распределение Фишера несимметрично относительно своего математического ожидания. Функция F-распределения:

$$Q(F, a, b) = I\left(\frac{b}{b + aF}, \frac{b}{2}, \frac{a}{2}\right),$$

где

$$I(x, u, v) = \frac{1}{\mathrm{B}(u, v)}\int_0^x t^{u-1}(1-t)^{v-1}dt$$

– неполная бета-функция.

Вероятность критического события находится различным образом в зависимости от того, относительно какой из альтернатив (*односторонней* или *двусторонней*) проверяется нулевая гипотеза. Если проверка выполняется относительно двусторонней альтернативы

$H_1 : \sigma_x^2 \neq \sigma_y^2$, то вероятность критического события находят по формуле

$$P(A) = \begin{cases} P', P' \le 1 \\ 2 - P', P' > 1 \end{cases},$$

где

$$P' = 2Q\left(\frac{\sigma_x^2}{\sigma_y^2}, N_1 - 1, N_2 - 1\right).$$

5.5. *Использование средств MS Excel для проверки гипотезы о равенстве средних: случай известных равных генеральных дисперсий*

Пусть из генеральных совокупностей с *известными* стандартными отклонениями $\sigma_x = \sigma_y = 10$ извлечены выборки объемом $N_1 = N_2 = 100$.

Пусть стандартные отклонения помещены в ячейки `A1` и `B1`, а варианты заполняют строки со второй по сто первую (ячейки `A2:A101` и `B2:B101`, соответственно).

	A	B
1	10	10
2	96,99768	110,8106
3	87,22317	103,1786
4	102,4426	97,91674
5	112,7647	95,4788
6	111,9835	98,36867
95	89,73069	107,2799
96	112,382	93,88923
97	96,88787	96,89761
98	91,60078	134,1644
99	91,78872	105,6607
100	95,71007	104,7085
101	95,46638	100,0033

Найдем выборочные средние. В ячейку `C1` введем формулу

`=СУММ(A2:A101)/100`

и переместим маркер автозаполнения до ячейки `D1`.

	A	B	C	D
1	10	10	=СУММ(A2:A101)/100	
2	96,99768	110,8106		

Вычислим значение связанной с разностью средних статистики, подчиненной стандартному нормальному распределению. В ячейку `G3` введем

`=(C1-D1)/(A1^2/100+B1^2/100)^0,5`

	A	B	C	D	E	F	G	H	I
1	10	10	99,59515	102,9814					
2	96,99768	110,8106							
3	87,22317	103,1786				Статистика:	=(C1-D1)/(A1^2/100+B1^2/100)^0,5		

Вероятность критического события можно найти, воспользовавшись функцией рабочего листа `НОРМСТРАСП`, возвращающей функцию распределения стандартного нормального закона. В ячейку `G4` введем

`=2-2*НОРМСТРАСП(ABS(G3))`

	A	B	C	D	E	F	G	H	I
1	10	10	99,59515	102,9814					
2	96,99768	110,8106							
3	87,22317	103,1786				Статистика:	-2,39441		
4	102,4426	97,91674				Вероятность критического события:	=2-2*НОРМСТРАСП(ABS(G3))		

Окончательный результат:

	A	B	C	D	E	F	G
1	10	10	99,59515	102,9814			
2	96,99768	110,8106					
3	87,22317	103,1786				Статистика:	-2,39441
4	102,4426	97,91674	Вероятность критического события:				0,016647

В данном примере вероятность критического события

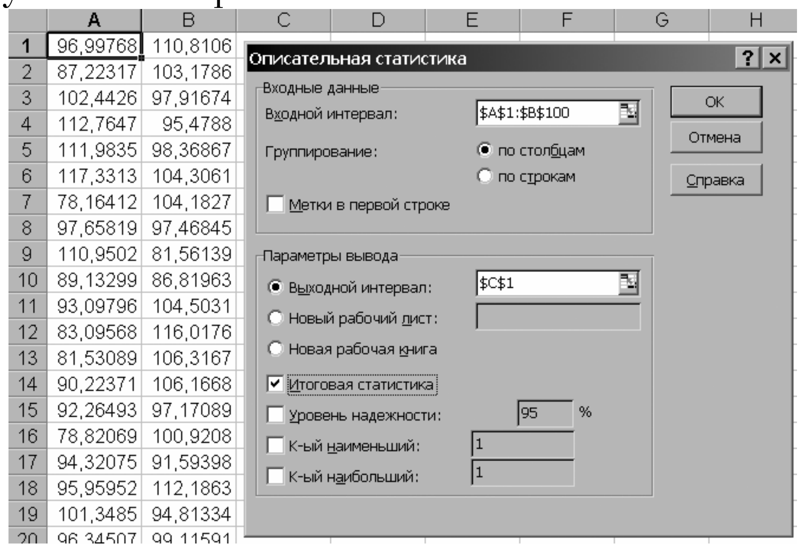
$$P(A) \approx 0{,}017 < 0{,}05,$$

поэтому на уровне значимости `0,05` гипотеза о равенстве средних должна быть отвергнута.

5.6. Использование средств MS Excel для проверки гипотезы о равенстве средних: случай неизвестных равных генеральных дисперсий

Пусть из генеральных совокупностей выборки объемом $N_1 = N_2 = 100$. Пусть варианты заполняют строки с первой по сотую (ячейки `A1:A100` и `B1:B100`, соответственно).

Для нахождения выборочных средних и выборочных дисперсий удобнее воспользоваться пакетом анализа. Из меню **Сервис** следует выбрать **Анализ данных**, далее – **Описательная статистика**. В качестве входного интервала следует указать два первых столбца (ячейки `$A$1:$B$100`). Результаты анализа можно поместить начиная с ячейки `C1`; установить флажок **Описательная статистика**.



Найдем смешанную оценку

$$D[\bar{x} - \bar{y}] = \frac{1}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\sum_{i=1}^{N_1}(x_i - \bar{x})^2 + \sum_{i=1}^{N_2}(y_i - \bar{y})^2\right)$$

дисперсии разности выборочных средних. В ячейку `G16` введем

```
=1/(D15+F15-2)*(1/D15+1/F15)*((D15-1)*D8+(F15-1)*F8)
```

8	97,65819	97,46845	Дисперси	117,8588	Дисперси	98,5838	
9	110,9502	81,56139	Эксцесс	-0,47571	Эксцесс	0,514209	
10	89,13299	86,81963	Асимметр	0,090701	Асимметр	0,182492	
11	93,09796	104,5031	Интервал	49,53235	Интервал	54,74467	
12	83,09568	116,0176	Минимум	74,22419	Минимум	79,41971	
13	81,53089	106,3167	Максимум	123,7565	Максимум	134,1644	
14	90,22371	106,1668	Сумма	9959,515	Сумма	10298,14	
15	92,26493	97,17089	Счет	100	Счет	100	
16	78,82069	100,9208				Смешанная оценка дисперсии:	=1/(D15+F15-2)*(1/D15+1/F15)*((D15-1)*D8+(F15-1)*F8)

Для вычисления статистики

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{D[\overline{x} - \overline{y}]}}$$

в ячейку G17 введем

```
=(D3-F3)/G16^0,5
```

	A	B	C	D	E	F	G	H
1	96,99768	110,8106	*Столбец1*		*Столбец2*			
2	87,22317	103,1786						
3	102,4426	97,91674	Среднее	99,59515	Среднее	102,9814		
4	112,7647	95,4788	Стандартı	1,085628	Стандартı	0,992894		
5	111,9835	98,36867	Медиана	99,15099	Медиана	102,7777		
6	117,3313	104,3061	Мода	#Н/Д	Мода	#Н/Д		
7	78,16412	104,1827	Стандартı	10,85628	Стандартı	9,928938		
8	97,65819	97,46845	Дисперси	117,8588	Дисперси	98,5838		
9	110,9502	81,56139	Эксцесс	-0,47571	Эксцесс	0,514209		
10	89,13299	86,81963	Асимметр	0,090701	Асимметр	0,182492		
11	93,09796	104,5031	Интервал	49,53235	Интервал	54,74467		
12	83,09568	116,0176	Минимум	74,22419	Минимум	79,41971		
13	81,53089	106,3167	Максимум	123,7565	Максимум	134,1644		
14	90,22371	106,1668	Сумма	9959,515	Сумма	10298,14		
15	92,26493	97,17089	Счет	100	Счет	100		
16	78,82069	100,9208				Смешанная оценка дисперсии:	2,164426	
17	94,32075	91,59398				Статистика:	=(D3-F3)/G16^0,5	

Вычислим вероятность критического события. В ячейку G18 введем

```
=СТЬЮДРАСП(ABS(G17);D15+F15-2;2)
```

	A	B	C	D	E	F	G	H	I	J
1	96,99768	110,8106	*Столбец1*		*Столбец2*					
2	87,22317	103,1786								
3	102,4426	97,91674	Среднее	99,59515	Среднее	102,9814				
4	112,7647	95,4788	Стандарті	1,085628	Стандарті	0,992894				
5	111,9835	98,36867	Медиана	99,15099	Медиана	102,7777				
6	117,3313	104,3061	Мода	#Н/Д	Мода	#Н/Д				
7	78,16412	104,1827	Стандарті	10,85628	Стандарті	9,928938				
8	97,65819	97,46845	Дисперси	117,8588	Дисперси	98,5838				
9	110,9502	81,56139	Эксцесс	-0,47571	Эксцесс	0,514209				
10	89,13299	86,81963	Асимметр	0,090701	Асимметр	0,182492				
11	93,09796	104,5031	Интервал	49,53235	Интервал	54,74467				
12	83,09568	116,0176	Минимум	74,22419	Минимум	79,41971				
13	81,53089	106,3167	Максимум	123,7565	Максимум	134,1644				
14	90,22371	106,1668	Сумма	9959,515	Сумма	10298,14				
15	92,26493	97,17089	Счет	100	Счет	100				
16	78,82069	100,9208				Смешанная оценка дисперсии:	2,164426			
17	94,32075	91,59398				Статистика:	-2,30167			
18	95,95952	112,1863				Вероятность критического события:	=СТЬЮДРАСП(ABS(G17);D15+F15-2;2)			

Окончательный результат:

16	78,82069	100,9208	Смешанная оценка дисперсии:	2,164426
17	94,32075	91,59398	Статистика:	-2,30167
18	95,95952	112,1863	Вероятность критического события:	0,022394

Вероятность критического события

$$P(A) \approx 0{,}022 < 0{,}05,$$

и на уровне значимости 0,05 гипотеза о равенстве средних должна быть отвергнута.

5.7. Использование средств MS Excel для проверки гипотезы о равенстве генеральных дисперсий

Пусть требуется сравнить дисперсии двух генеральных совокупностей на основе извлеченных выборок объемами $N_1 = N_2 = 20$. Пусть варианты помещены в первые два столбца и заполняют строки со второй по двадцатую (ячейки A2:A20 и B2:B101, соответственно).

	A	B
1	-93,8348	75,51458
2	-115,126	97,10953
3	-92,3503	87,93137
4	-91,7102	103,9361
5	-88,7351	98,03636
6	-82,0142	108,3103
7	-111,252	102,4355
8	-97,1876	78,76266
9	-92,9408	81,49287
10	-101,148	98,96504
11	-94,6158	96,11427
12	-109,395	113,1854
13	-84,5413	109,8481
14	-111,39	111,3724
15	-97,1629	110,9127
16	-125,431	96,70991
17	-91,6994	103,162
18	-100,871	96,92318
19	-83,4188	100,4649
20	-95,8188	101,1612

Для нахождения оценок дисперсий воспользуемся средствами пакета анализа (аналогично предыдущей задаче). Из меню Сервис следует выбрать Анализ данных, далее – Описательная статистика. В качестве входного интервала следует указать два первых столбца (ячейки $A$1:$B$20). Результаты анализа можно поместить начиная с ячейки C1; установить флажок Описательная статистика.

Вычислим подчиненную F-распределению статистику. В ячейку G16 поместим формулу

=D8/F8

8	-97,1876	78,76266	Дисперси	129,2361	Дисперси	114,2354	
9	-92,9408	81,49287	Эксцесс	0,291126	Эксцесс	0,114196	
10	-101,148	98,96504	Асимметр	-0,80963	Асимметр	-0,80253	
11	-94,6158	96,11427	Интервал	43,41709	Интервал	37,67082	
12	-109,395	113,1854	Минимум	-125,431	Минимум	75,51458	
13	-84,5413	109,8481	Максимум	-82,0142	Максимум	113,1854	
14	-111,39	111,3724	Сумма	-1960,64	Сумма	1972,348	
15	-97,1629	110,9127	Счет	20	Счет	20	
16	-125,431	96,70991				Статистика:	=D8/F8

Для вычисления величины

$$P' = 2Q\left(\frac{\sigma_x^2}{\sigma_y^2}, N_1 - 1, N_2 - 1\right)$$

в ячейку G17 поместим формулу

=FРАСП(G16;D15-1;F15-1)

14	-111,39	111,3724	Сумма	-1960,64	Сумма	1972,348	
15	-97,1629	110,9127	Счет	20	Счет	20	
16	-125,431	96,70991				Статистика:	1,131314
17	-91,6994	103,162		Односторонная вероятность:			=FРАСП(G16;D15-1;F15-1)

Наконец, для нахождения вероятности критического события в ячейку G18 введем

=ЕСЛИ(G17<1;G17;2-G17)

16	-125,431	96,70991		Статистика:	1,131314
17	-91,6994	103,162		Односторонная вероятность:	0,395388
18	-100,871	96,92318	Вероятность критического события:		=ЕСЛИ(G17<1;G17;2-G17)

Окончательный результат:

	A	B	C	D	E	F	G
1	-93,8348	75,51458	*Столбец1*		*Столбец2*		
2	-115,126	97,10953					
3	-92,3503	87,93137	Среднее	-98,0321	Среднее	98,61741	
4	-91,7102	103,9361	Стандарті	2,542008	Стандарті	2,389931	
5	-88,7351	98,03636	Медиана	-95,2173	Медиана	99,71496	
6	-82,0142	108,3103	Мода	#Н/Д	Мода	#Н/Д	
7	-111,252	102,4355	Стандарті	11,36821	Стандарті	10,68809	
8	-97,1876	78,76266	Дисперси	129,2361	Дисперси	114,2354	
9	-92,9408	81,49287	Эксцесс	0,291126	Эксцесс	0,114196	
10	-101,148	98,96504	Асимметр	-0,80963	Асимметр	-0,80253	
11	-94,6158	96,11427	Интервал	43,41709	Интервал	37,67082	
12	-109,395	113,1854	Минимум	-125,431	Минимум	75,51458	
13	-84,5413	109,8481	Максимум	-82,0142	Максимум	113,1854	
14	-111,39	111,3724	Сумма	-1960,64	Сумма	1972,348	
15	-97,1629	110,9127	Счет	20	Счет	20	
16	-125,431	96,70991				Статистика:	1,131314
17	-91,6994	103,162			Односторонная вероятность:		0,395388
18	-100,871	96,92318		Вероятность критического события:			0,395388

В данном примере $P(A) \approx 0{,}4 > 0{,}05$, поэтому на уровне значимости $\alpha = 0{,}05$ нет оснований отвергать гипотезу о равенстве дисперсий в пользу двусторонней альтернативы.

6. Задачи регрессионного анализа и математической теории эксперимента

В рамках регрессионного анализа объединяются задачи, связанные с *построением функциональныхязависимостей* между двумя или несколькими числовыми переменными. Регрессионный анализ является основным средством концентрации, «свертки» эмпирической информации. Подобные операции иногда упрощенно называют *сглаживанием экспериментальныэ данныэ*. Следует помнить, что для многих задач регрессионного анализа характерна более широкая постановка, включающая статистический анализ полученных результатов.

Встречающиеся на практике системы можно считать детерминированными, однако число составляющих в них велико. Поэтому свойства таких систем могут быть исследованы только статистическими методами, а анализ всегда базируется на вероятностных представлениях.

Пусть над некоторой системой производится эксперимент, в ходе которого имеется возможность произвольно выбирать – *варьировать* – значения n независимых входных переменных

$$(x_1, x_2, \ldots, x_n) = \mathbf{x}$$

– *варьируемыэ - акторов*.

Исследуемая система может пониматься как *черный ящик* (рис. 6.1). Внутреннее содержание системы остается неизвестным для исследователя. Регистрации доступны лишь значения зависимой переменной – *откликая системы*. На измеренное значение отклика оказывают влияние как объективные закономерности функционирования системы, так и случайные факторы. Последние выражают либо внутренне присущую отклику изменчивость, либо влияние на него обстоятельств, не

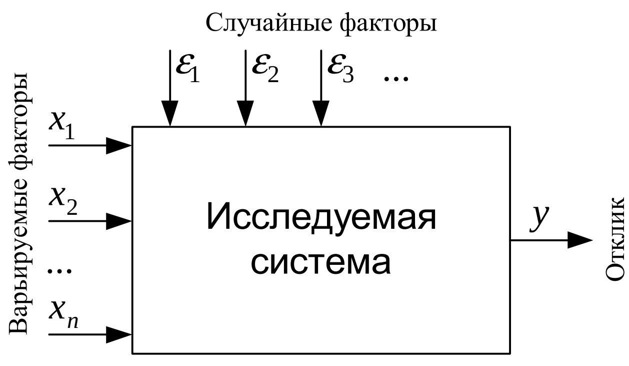


Рис. 6.1. Исследуемая система – «черный ящик»

учтенных в эксперименте (в частности, они могут выражать и влияние несовершенства средств измерений). Путь при отсутствии случайных факторов связь между откликом и входными переменными дается зависимостью $y = \psi(\mathbf{x})$. Тогда наблюдаемое значение отклика можно представить в виде суммы

$$y = \psi(\mathbf{x}) + \varepsilon,$$

в которой первое слагаемое закономерно зависит от $\mathbf{x}$, а второе связано с влиянием случайных факторов. Это слагаемое условно можно назвать «ошибкой» эксперимента.

При обработке эмпирического материала возникает необходимость восстановления аналитической (функциональной) зависимости отклика от варьируемых факторов:

$$y = f(x_1, x_2, \ldots, x_n) = f(\mathbf{x})$$

– *экспериментально-статистической* (ЭС) *модели*, которая являлась бы в некотором смысле наилучшим описанием исследуемой системы. Сразу оговоримся, что изложение пока не предполагает ни *повторенияяопытов*, ни решения вопроса об *адекватности* полученного описания – соответствия его (неизвестной) истинной зависимости $y = \psi(\mathbf{x})$.

Как правило, общий вид модели (вид аналитической зависимости) выбирается заранее. Выбор модели является неформальной операцией и определяется основанной на накопленном опыте интуицией исследователя и доступной информацией об объекте исследования.

В простейших случаях выбор модели можно производить только на основе эмпирических данных. Например, если полученные при N измерениях пары значений (x_u, y_u), $u = \overline{1, N}$ скалярного варьируемого фактора x и отклика y сгруппированы вблизи прямой линии, то в качестве модели можно выбрать линейную функцию $y = b_0 + b_1 x$; если значения сгруппированы вблизи параболы, то можно взять квадратичную модель $y = b_0 + b_1 x + b_{11} x^2$, и т.д.

Очевидно, что для «хорошей» модели предсказанное значение $f(\mathbf{x})$ в точке $\mathbf{x}$ должно быть по возможности «близко» к наблюдаемому в эксперименте значению отклика. Вопрос о степени «близости» допускает различную трактовку. Обычно для сравнения предсказанных и эмпирических значений определяется *целевая функция*, которая зависит от различия между опытными и предсказанными значениями. После введения целевой функции задача обработки эмпирического материала сводится к поиску экстремума (обычно – минимума) целевой функции и может решаться известными средствами математического анализа.

В общем случае модель

$$y = f\left(\mathbf{x}, b_1, b_2, \dots, b_L\right) \tag{6.1}$$

включает L неизвестных параметров b_1, b_2, …, b_L. Их значения выбираются так, чтобы целевая функция (от L переменных – неизвестных параметров модели), характеризующая различие между предсказанными и наблюдаемыми значениями отклика, достигла минимума.

Как уже было отмечено, выбор общего вида модели – неформальная операция, для которой ни теория вероятностей, ни математическая статистика не предоставляют никаких средств. В то же время выбор подлежащей минимизации целевой функции, связанной с «близостью» предсказанных и эмпирических значений, при некоторых предположениях может быть сделан формально. Основой выбора является *принципямаксимальногояправдоподобия*, согласно которому *наилучшим описанием исследуемой системы вл - етсяятакое,япр, якоторомямакс, мальнаявероятностьяполуч, тья, з- меренныеянаяопытеязначен, яяоткл, ка*.

Пусть истинная (неизвестная исследователю) зависимость отклика от варьируемых факторов дается выражением $y = \psi(\mathbf{x})$. Предположим, что влияющие на систему случайные факторы таковы, что выполнены три условия.

1. Ошибки измерений отклика (разности ε между эмпирическим y_u и неизвестным истинным $\psi(\mathbf{x}_u)$ значениями в u-м опыте) *распределеныяпояномальномуязакону*; математическое ожидание отклика при этом оказывается равным истинному значению $\psi(\mathbf{x}_u)$.

2. Измерения независимы и *равноточны* – стандартные отклонения отклика (и ошибок измерений) во всех опытах постоянны и равны σ.

Вместе с принципом максимального правдоподобия сделанные предположения составляют основу большинства методов регрессионного анализа. При этих предположениях в каждом u-м из N опытов результат измерения y_u будет случайной величиной Y_u, плотность вероятности которой

$$g_u(y_u)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{[y_u-\psi(\mathbf{x}_u)]^2}{2\sigma^2}\right).$$

Эксперимент есть событие, состоящее в том, что случайные величины Y_1, Y_2, …, Y_N приняли значения y_1, y_2, …, y_N. Так как случайные величины Y_u непрерывны, то вероятность события $Y_u = u_u$ равна нулю; можно лишь поставить вопрос о вероятности попадания величины Y_u на малый интервал $[y_u, y_u+dy_u)$. Эта вероятность

$$P(y_u \le Y_u < y_u+dy_u)=g_u(y_u)dy_u=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{[y_u-\psi(\mathbf{x}_u)]^2}{2\sigma^2}\right)dy_u.$$

Так как измерения выполняются независимо, то вероятность P произведения событий $Y_u\in[y_u, y_u+dy_u)$, $u=\overline{1,N}$ оказывается равной произведению вероятностей сомножителей:

$$P=\prod_{u=1}^{N}P(y_u \le Y_u < y_u+dy_u)=\prod_{u=1}^{N}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{[y_u-\psi(\mathbf{x}_u)]^2}{2\sigma^2}\right)dy_u=$$

$$=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N\exp\left(-\frac{1}{2\sigma^2}\sum_{u=1}^{N}[y_u-\psi(\mathbf{x}_u)]^2\right)dy_1dy_2...dy_n,$$

Обозначим

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N dy_1dy_2...dy_n=K$$

– величина, не зависящая ни от номера u опыта, ни от характера зависимости $\psi(\mathbf{x}_u)$ отклика от входных переменных. Тогда вероятность получить значения, близкие к наблюдаемым на опыте:

$$P=K\exp\left(-\frac{1}{2\sigma^2}\sum_{u=1}^{N}[y_u-\psi(\mathbf{x}_u)]^2\right).$$

Эта вероятность возрастает вместе с увеличением показателя степени и максимальна, если сумма в показателе достигает *минимума*:

$$\sum_{u=1}^{N}[y_u-\psi(\mathbf{x}_u)]^2=\min.$$

Но функция $\psi(\mathbf{x})$ неизвестна, а задача исследователя как раз и состоит в нахождении модели $f(\mathbf{x})$, которая была бы близка к ней.

Поэтому при сделанных предположениях из принципа максимального правдоподобия следует, что *наилучшей моделью будетятакая,я дляя которойя суммая квадратовя отклонен, йя хмп, р, ческ, хя значен, й* y_u *отязначен, й,япредсказанныхямоделью,яобращаетсяявя м, н, мум*:

$$S(y_u, f, \mathbf{x}_u) = \sum_{u=1}^{N} \left[y_u - f(\mathbf{x}_u) \right]^2 = \min . \qquad (6.2)$$

Методомя на, меньш, хя квадратов называют метод отыскания модели, который обеспечивает выполнение указанного условия.

7. Подбор параметров линейной модели

Пусть в процессе исследования варьировалась одна независимая переменная x и после проведения N экспериментов получены значения y_u, $u = \overline{1, N}$. Требуется методом наименьших квадратов подобрать параметры линейной экспериментально-статистической модели

$$y = ax + b .$$

Для модели указанного вида сумма квадратов отклонений имеет вид

$$S = \sum_{u=1}^{N} \left(y_u - (ax_u + b) \right)^2 .$$

Считая эту сумму функцией неизвестных параметров

$$S = S(a, b),$$

потребуем выполнения необходимого условия локального экстремума

$$\begin{cases} \dfrac{\partial S}{\partial a} = 0 \\ \dfrac{\partial S}{\partial b} = 0 \end{cases} .$$

Дифференцируя и приравнивая частные производные к нулю, получим:

$$\begin{cases} \sum\limits_{u=1}^{N} (y_u - ax_u - b) x_u = 0 \\ \sum\limits_{u=1}^{N} (y_u - ax_u - b) = 0 \end{cases} .$$

Изменим порядок суммирования:

$$\begin{cases} a\sum\limits_{u=1}^{N} x_u^2 + b\sum\limits_{u=1}^{N} x_u = \sum\limits_{u=1}^{N} y_u x_u \\ a\sum\limits_{u=1}^{N} x_u + bN = \sum\limits_{u=1}^{N} y_u \end{cases} . \tag{7.1}$$

Если все значения x_u различны, то полученная система двух линейных уравнений (которую называют *нормальной системой*) имеет единственное решение. Можно доказать, что это решение действительно соответствует точке локального минимума функции $S = S(a,b)$.

8. Случай модели, линейной по параметрам

Нормальная система, возникающая в процессе применения метода наименьших квадратов, определяется видом экспериментально-статистической модели $y = f(\mathbf{x})$. В большинстве случаев нормальная система является нелинейной и решается только численно.

Однако существует достаточно широкий класс практически важных моделей, для которых нормальная система является линейной и допускает простое и компактное представление. Этот класс представлен *моделям, ,ял, нейным, япоя параметрам*:

$$f(\mathbf{x}) = b_1\varphi_1(\mathbf{x}) + b_2\varphi_2(\mathbf{x}) + \ldots + b_L\varphi_L(\mathbf{x}).$$

В этих моделях функции $\varphi_j(\mathbf{x})$, $j = \overline{1,L}$ носят название *баз, сныхя функций*. Примерами моделей, линейных по параметрам, являются модели

$$y = ax + b \text{ (базисные функции: } \varphi_1(x) = x,\ \varphi_2(x) = 1);$$

$$y = ax^2 + b\sin x + ce^x \ (\varphi_1(x) = x^2, \varphi_2(x) = \sin x, \varphi_3(x) = e^x);$$

$$z = b_0 + b_1x + b_2y + b_{11}x^2 + b_{12}xy + b_{22}y^2$$

$$(\varphi_1 = 1,\ \varphi_2 = x,\ \varphi_3 = y,\ \varphi_4 = x^2,\ \varphi_5 = xy,\ \varphi_6 = y^2) \text{ и т.д.}$$

Для моделей указанного вида

$$f(\mathbf{x}) = \sum_{j=1}^{L} b_j\varphi_j(\mathbf{x})$$

сумма квадратов отклонений предсказанных $f(\mathbf{x}_u)$ и экспериментальных значений y_u имеет вид

$$S(b_1, b_2, \ldots, b_L) = \sum_{u=1}^{N} \left(y_u - f(\mathbf{x}_u)\right)^2 = \sum_{u=1}^{N} \left(y_u - \sum_{j=1}^{L} b_j\varphi_j(\mathbf{x}_u) \right)^2 .$$

Считая эту сумму функцией от L неизвестных параметров, потребуем выполнения необходимого условия локального экстремума

$$\sum_{u=1}^{N}\left[\left(y_u-\sum_{j=1}^{L}b_j\varphi_j(\mathbf{x}_u)\right)\left(\frac{\partial}{\partial b_i}\sum_{j=1}^{L}b_j\varphi_j(\mathbf{x}_u)\right)\right]=0,\ i=\overline{1,L}.$$

Базисные функции не зависят от параметров, поэтому при всех $i\neq j$ производные

$$\frac{\partial}{\partial b_i}b_j\varphi_j(\mathbf{x}_u)$$

равны нулю. Следовательно, в каждой из сумм $\frac{\partial}{\partial b_i}\sum_{j=1}^{L}b_j\varphi_j(\mathbf{x}_u)$ имеется только одно ненулевое слагаемое:

$$\frac{\partial}{\partial b_i}\sum_{j=1}^{L}b_j\varphi_j(\mathbf{x}_u)=\varphi_i(\mathbf{x}_u).$$

Нормальная система принимает вид

$$\sum_{u=1}^{N}\left[\left(y_u-\sum_{j=1}^{L}b_j\varphi_j(\mathbf{x}_u)\right)\varphi_i(\mathbf{x}_u)\right]=0,\ i=\overline{1,L}.$$

Изменим порядок суммирования и перенесем в правую часть слагаемые, в которые входят эмпирические значения отклика:

$$\sum_{u=1}^{N}\left(\varphi_i(\mathbf{x}_u)\sum_{j=1}^{L}b_j\varphi_j(\mathbf{x}_u)\right)=\sum_{u=1}^{N}y_u\varphi_i(\mathbf{x}_u),\ i=\overline{1,L}.$$

Вновь меняя в левой части порядок суммирования, запишем нормальную систему в виде

$$\sum_{j=1}^{L}\left(b_j\sum_{u=1}^{N}\varphi_i(\mathbf{x}_u)\varphi_j(\mathbf{x}_u)\right)=\sum_{u=1}^{N}y_u\varphi_i(\mathbf{x}_u),\ i=\overline{1,L}. \qquad (8.1)$$

Введем обозначения:

$$\mathbf{X}=\begin{pmatrix}\varphi_1(\mathbf{x}_1) & \varphi_2(\mathbf{x}_1) & \dots & \varphi_L(\mathbf{x}_1)\\ \varphi_1(\mathbf{x}_2) & \varphi_2(\mathbf{x}_2) & \dots & \varphi_L(\mathbf{x}_2)\\ \dots & & & \\ \varphi_1(\mathbf{x}_N) & \varphi_2(\mathbf{x}_N) & \dots & \varphi_L(\mathbf{x}_N)\end{pmatrix} -$$

матрица размера NxL, называемая *матрицей базисных функций*;

$$\mathbf{B}=(b_1,b_2,\dots,b_L)^T -$$

вектор - столбец высоты $й$, называемый *вектором искомых параметров*;

$$\mathbf{Y} = \left(y_1, y_2, \ldots, y_N\right)^T \ -$$

вектор - столбец высоты N, называемый *вектором откликов*.

Тогда нормальную систему (8.1) можно записать в матричной форме:

$$\left(\mathbf{X}^T \mathbf{X}\right)\mathbf{B} = \mathbf{X}^T \mathbf{Y}. \qquad (8.2)$$

Входящая в эту систему симметрическая квадратная матрица

$$\mathbf{M} = \mathbf{X}^T \mathbf{X} = \begin{pmatrix} \varphi_1(\mathbf{x}_1) & \varphi_1(\mathbf{x}_2) & \ldots & \varphi_1(\mathbf{x}_N) \\ \varphi_2(\mathbf{x}_1) & \varphi_2(\mathbf{x}_2) & \ldots & \varphi_2(\mathbf{x}_N) \\ \ldots & & & \\ \varphi_L(\mathbf{x}_1) & \varphi_L(\mathbf{x}_2) & \ldots & \varphi_L(\mathbf{x}_N) \end{pmatrix} \begin{pmatrix} \varphi_1(\mathbf{x}_1) & \varphi_2(\mathbf{x}_1) & \ldots & \varphi_L(\mathbf{x}_1) \\ \varphi_1(\mathbf{x}_2) & \varphi_2(\mathbf{x}_2) & \ldots & \varphi_L(\mathbf{x}_2) \\ \ldots & & & \\ \varphi_1(\mathbf{x}_N) & \varphi_2(\mathbf{x}_N) & \ldots & \varphi_L(\mathbf{x}_N) \end{pmatrix} =$$

$$= \begin{pmatrix} \sum\limits_{u=1}^{N} \varphi_1^2(\mathbf{x}_u) & \sum\limits_{u=1}^{N} \varphi_1(\mathbf{x}_u)\varphi_2(\mathbf{x}_u) & \ldots & \sum\limits_{u=1}^{N} \varphi_1(\mathbf{x}_u)\varphi_L(\mathbf{x}_u) \\ \sum\limits_{u=1}^{N} \varphi_2(\mathbf{x}_u)\varphi_1(\mathbf{x}_u) & \sum\limits_{u=1}^{N} \varphi_2^2(\mathbf{x}_u) & \ldots & \sum\limits_{u=1}^{N} \varphi_2(\mathbf{x}_u)\varphi_L(\mathbf{x}_u) \\ \ldots & & & \\ \sum\limits_{u=1}^{N} \varphi_L(\mathbf{x}_u)\varphi_1(\mathbf{x}_u) & \sum\limits_{u=1}^{N} \varphi_L(\mathbf{x}_u)\varphi_2(\mathbf{x}_u) & \ldots & \sum\limits_{u=1}^{N} \varphi_L^2(\mathbf{x}_u) \end{pmatrix},$$

порядок которой совпадает с числом базисных функций (и с числом слагаемых в экспериментально-статистической модели), называется *матрицей моментов*, или *информационнойяматрицей*[1].

Рассмотренная выше одномерная линейная регрессия

$$y = b + ax$$

является частным случаем модели, линейной по параметрам. Для этой модели базисные функции

$$\varphi_1(x) = 1, \ \varphi_2(x) = x,$$

поэтому ее матрица базисных функций (имеющая N строк и два столбца):

[1] Согласно ГОСТ 24026-80, информационная матрица есть *частноеяотяделенияяматрицыямо-ментовяналячислояопытовя*N. Однако в некоторых источниках понятия информационной матрицы и матрицы моментов отождествляют, а матрицу $\mathbf{X}^T\mathbf{X}/N$ называют *нормированнойяин-- ормационнойяматрицей*.

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_N \end{pmatrix}.$$

Матрица моментов одномерной линейной регрессии:

$$\mathbf{M} = \mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \\ 1 & x_N \end{pmatrix} = \begin{pmatrix} N & \sum_{u=1}^{N} x_u \\ \sum_{u=1}^{N} x_u & \sum_{u=1}^{N} x_u^2 \end{pmatrix};$$

нетрудно видеть, что элементами этой матрицы являются именно те суммы, которые входят в левые части нормальной системы (7.1).

Матрица, обратная к матрице моментов,

$$\mathbf{D} = \mathbf{M}^{-1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

называется *ковариационной матрицей*, или *матрицей ошибок*. Искомый столбец параметров равен

$$\mathbf{B} = \mathbf{D}\mathbf{X}^T\mathbf{Y}. \tag{8.3}$$

Применение соотношения (8.3) приводит к L-кратному увеличению вычислительных затрат по сравнению с (8.2). Несмотря на это для нахождения параметров целесообразно использовать именно соотношение (8.3). Это связано с тем, что диагональные элементы ковариационной матрицы характеризуют дисперсии параметров модели, а внедиагональные – их «взаимное влияние»[1]. Ковариационная матрица требуется на этапе статистического анализа построенной ЭС-модели.

Для рассмотренной одномерной линейной регрессии ковариационная матрица равна

$$\mathbf{D} = \begin{pmatrix} N & \sum_{u=1}^{N} x_u \\ \sum_{u=1}^{N} x_u & \sum_{u=1}^{N} x_u^2 \end{pmatrix}^{-1} = \frac{1}{N\sum_{u=1}^{N} x_u^2 - \sum_{u=1}^{N} x_u} \begin{pmatrix} \sum_{u=1}^{N} x_u^2 & -\sum_{u=1}^{N} x_u \\ -\sum_{u=1}^{N} x_u & N \end{pmatrix}.$$

[1] Строгое изложение соответствующих понятий приведено ниже.

Взаимное влияние параметров будет минимальным, если внедиагональные элементы обратятся в ноль; получаем условие

$$\sum_{u=1}^{N} x_u = 0\,. \tag{8.4}$$

При проведении *активного эксперимента* исследователь почти всегда имеет возможность выбрать значения входных переменных так, чтобы обеспечить выполнение условий, подобных (8.4). В этом случае говорят, что *план эксперимента* в том или ином смысле *оптимален*.

Если для линейной регрессионной модели условие (8.4) выполнено, то ее ковариационная матрица

$$\mathbf{D} = \frac{1}{N\sum_{u=1}^{N} x_u^2}\begin{pmatrix} \sum_{u=1}^{N} x_u^2 & 0 \\ 0 & N \end{pmatrix} = \begin{pmatrix} \dfrac{1}{N} & 0 \\ 0 & \left(\sum_{u=1}^{N} x_u^2\right)^{-1} \end{pmatrix},$$

поэтому дисперсии параметров модели пропорциональны величинам

$$D_a = \left(\sum_{u=1}^{N} x_u^2\right)^{-1}, \;\; D_b = \frac{1}{N}\,.$$

Как и следовало ожидать, дисперсия свободного члена b обратно пропорциональна числу опытов. Дисперсия коэффициента a при входной переменной уменьшается вместе с возрастанием числа опытов и увеличением абсолютных величин тех значений входной переменной, для которых измеряются значения отклика.

8.1. Использование средств MS Excel для построения одномерной линейной регрессионной модели

Пусть для четырех выбранных значений x_u, $u = \overline{1{,}4}$ независимой переменной x выполнен эксперимент и получены значения отклика y_u (табл. 8.1).

Таблица 8.1

Эмпирические значения отклика

	Номер эксперимента			
	1	2	3	4
Значение переменной x	0	1	2	3
Значение отклика y	2	2	3	4

Пусть значения входной переменной заполняют первые четыре ячейки первого столбца (`A1:A4`), а значения отклика заполняют аналогичные ячейки второго столбца (`B1:B4`).

Здесь мы намеренно *откажемся* от использования общей формулы (8.3) для параметров модели.

Отдельно вычислим суммы

$$\sum_{u=1}^{N} x_u\,,\ \sum_{u=1}^{N} x_u^2\,,\ \sum_{u=1}^{N} y_u\,,\ \sum_{u=1}^{N} x_u y_u\,.$$

Удобно заполнить диапазон `C1:C4` квадратами значений входной переменной, а диапазон `D1:D4` – произведениями значений входной переменной и отклика. Поместим в ячейку `C1` формулу

```
=A1^2
```

а в ячейку `D1` – формулу

```
=A1*B1
```

	A	B	C	D
1	0	2	=A1^2	0
2	1	2		
3	2	3		
4	3	4		

	A	B	C	D
1	0	2	0	=A1*B1
2	1	2		
3	2	3		
4	3	4		

После этого выделим диапазон `C1:D1` и переместим маркер автозаполнения до четвертой строки.

Для нахождения сумм, входящих в нормальную систему (7.1), поместим в ячейку `A5` формулу

```
=СУММ(A1:A4)
```

(можно выделить диапазон `A5:A1` и нажать на панели инструментов кнопку автосуммы Σ), после чего (при единственной выделенной ячейке `A5`) переместим маркер автозаполнения до четвертого столбца.

	A	B	C	D
1	0	2	0	0
2	1	2	1	2
3	2	3	4	6
4	3	4	9	12
5	=СУММ(A1:A4)			

	A	B	C	D	E
1	0	2	0	0	
2	1	2	1	2	
3	2	3	4	6	
4	3	4	9	12	
5	6	11	14	=СУММ(D1:D4)	

Нетрудно заметить, что условие (8.4) для выбранного *плана эксперимента* (значений входной переменной в табл. 8.1) не выполнено: сумма в ячейке `A5` отлична от нуля. Это говорит не в пользу приведенного примера.

Нормальная система имеет вид

$$\begin{cases} 14a + 6b = 20 \\ 6a + 4b = 11 \end{cases}.$$

Решение «придется» искать по формулам Крамера (вновь подчеркнем – соотношение (8.3) и эффективные вычислительные методы из группы методов Гаусса было решено не использовать). Для системы второго порядка это нетрудно сделать «вручную», однако Excel оказывается полезным и здесь.

	A	B	C	D
1	0	2	=A1^2	=A1*B1
2	1	2	=A2^2	=A2*B2
3	2	3	=A3^2	=A3*B3
4	3	4	=A4^2	=A4*B4
5	=СУММ(A1:A4)	=СУММ(B1:B4)	=СУММ(C1:C4)	=СУММ(D1:D4)
6				
7	=C5	=A5		
8	=B7	4		
9	Δ	**=МОПРЕД(A7:B8)**		
10	=D5	=B7		
11	=B5	=B8		
12	Δ1	**=МОПРЕД(A10:B11)**		
13	=A7	=A10		
14	=A8	=A11		
15	Δ2	**=МОПРЕД(A13:B14)**		

	A	B	C	D
1	0	2	0	0
2	1	2	1	2
3	2	3	4	6
4	3	4	9	12
5	6	11	14	20
6				
7	14	6		
8	6	4		
9	Δ	**20**		
10	20	6		
11	11	4		
12	Δ1	**14**		
13	14	20		
14	6	11		
15	Δ2	**34**		

Имеем:

$$a = \frac{14}{20} = 0{,}7\,;\ \ b = \frac{34}{20} = 1{,}7\,.$$

7	14	6		**a**	**=B12/B9**
8	6	4		**b**	**1,7**
9	Δ	**20**			
10	20	6			
11	11	4			
12	Δ1	**14**			
13	14	20			
14	6	11			
15	Δ2	**34**			

7	14	6		**a**	**0,7**
8	6	4		**b**	**=B15/B9**
9	Δ	**20**			
10	20	6			
11	11	4			
12	Δ1	**14**			
13	14	20			
14	6	11			
15	Δ2	**34**			

Искомая модель:

$$y = 0{,}7x + 1{,}7\,.$$

	A	B	C	D	E	F	G
1	0	2	0	0		=A1*$E$7+$E$8	
2	1	2	1	2		2,4	
3	2	3	4	6		3,1	
4	3	4	9	12		3,8	
5	6	11	14	20			
6							
7	14	6		**a**	**0,7**		
8	6	4		**b**	**1,7**		

Полезно построить диаграмму, на которой были бы отмечены как эмпирические значения, так и график предсказанного моделью отклика. Для построения диаграммы придется табулировать предсказанные значения (хотя в данном случае регрессия линейная, и хватило бы двух точек). Поместим в какую-либо ячейку первой строки (например – в F1) формулу

=A1*$E$7+$E$8
и переместим маркер автозаполнения до четвертой строки.

Выделим ячейки A1:C4, из меню Вставка выберем Диаграмма, далее – Точечная. На вкладке Ряд в списке Ряд для второго ряда данных следует изменить значения Y на
=’шаг 4’!$F$1:$F$4
где «шаг 4» – имя рабочего листа, на который помещается диаграмма. Вместо корректировки значений можно до вставки диаграммы выделить два диапазона: A1:B4 и F1:F4 (следует выделить первый из них, нажать и удерживать клавишу Ctrl, и при нажатой левой кнопке мыши переместить курсор от ячейки F1 до F4).

После вставки диаграммы для первого ряда следует отключить построение интерполяционной сплайновой кривой (двойной щелчок по любому элементу ряда, на вкладке Вид в группе Линия установить Отсутствует), а для второго включить ее построение (аналогично, но в группе Линия установить Обычная). Установка флажка Сглаженная линия, включающая сплайновую интерполяцию табулированных значений, в данном случае будет излишней, но для последующих примеров может оказаться полезной.

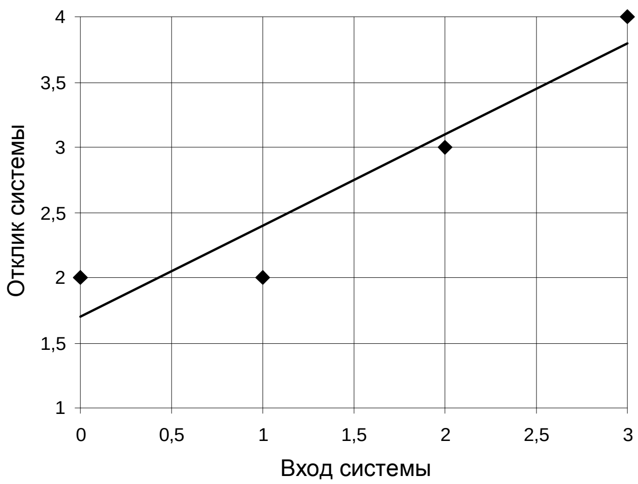


Рис. 8.1. Эмпирические значения и график предсказанного моделью отклика

9. Основные понятия математической теории эксперимента

Выше были рассмотрены исходные предпосылки построения ЭС-моделей. Исходя из весьма общих предположений о характере ошибок измерений, для вектора коэффициентов линейной по параметрам модели было получено соотношение

$$\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{M}^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{D}\mathbf{X}^T\mathbf{Y},$$

где $\mathbf{M} = \mathbf{X}^T\mathbf{X}$ – матрица моментов, $\mathbf{D} = \mathbf{M}^{-1}$ – ковариационная матрица.

Закономерно возникают вопросы: как влияет выбор значений входной переменной на оценки параметров и какими должны быть эти значения, чтобы оценки были в некотором смысле «наилучшими». Ответы на эти вопросы определяются структурой ковариационной матрицы (и, следовательно, зависят также от выбранной ЭС-модели).

В рамках регрессионного анализа определяются только методы поиска и последующей оценки статистической значимости параметров экспериментально-статистической модели. Вопрос об оптимальном выборе значений входной переменной решается в рамках *математической теории эксперимента* – дисциплины, основная цель методов которой заключается в извлечении максимального количества объективной информации о влиянии факторов на исследуемый процесс при помощи наименьшего числа наблюдений. Методы теории эксперимента широко используются для выбора оптимального состава многокомпонентных смесей, повышения производительности оборудования, повышения качества продукции.

Для дальнейшего изложения нам потребуется ряд определений.

Координатное пространство, переменными в котором являются варьируемые факторы X_1, X_2, ..., X_k, называют *факторнымяпроф странством*. Число k варьируемых факторов называют *размерноф стью* факторного пространства. Каждому из N экспериментов соответствует определенная точка факторного пространства; понятия «эксперимент» и «точка факторного пространства» и «экспериментальная точка» можно считать тождественными. График экспериментально-статистической модели – геометрический образ в $\left(k+1\right)$-мерном пространстве – называют *поверхностьюяотклика*.

Диапазон изменения входной переменной X_i, $i=\overline{1,N}$ называют *размахомяварьирования*, а половину этого диапазона

$$\Delta X_i = \frac{X_{i,\max} - X_{i,\min}}{2}$$

называют *интерваломяварьирования*. Значения входной переменной называют *уровнямияварьирования*. Середина диапазона варьирования

$$X_{i,0} = \frac{X_{i,\max} + X_{i,\min}}{2}$$

называется *основнымяуровнем* фактора.

Планомяхксперимента называют число и способ размещения экспериментальных точек в факторном пространстве. *Матрицейя плана* называют матрицу размера Nxk (где N – число экспериментов, k – размерность факторного пространства), в строках которой находятся координаты экспериментальных точек.

Удобство построения, статистического анализа и последующей интерпретации ЭС-моделей увеличивается при переходе от исходных – *натуральных* – действующих переменных к безразмерным *нормализованным* переменным

$$x_i = \frac{X_i - X_{i,0}}{\Delta X_i}.$$

Целесообразность этой операции с точки зрения простоты интерпретации модели иллюстрируется следующим примером. Пусть оценивается влияние двух факторов – времени тепловой обработки X_1 и массы[1] модификатора X_2 – на прочность R полимерной мастики. Каждый из факторов варьируется на двух уровнях. План эксперимента образован четырьмя точками (первые два столбца табл. 9.1).

Таблица 9.1

X_1, ч	X_2, кг	R, МПа
2	$1\ 10^{-5}$	121
2	$2\ 10^{-5}$	129
4	$1\ 10^{-5}$	148
4	$2\ 10^{-5}$	154

Для прочности выбрана линейная двухфакторная ЭС-модель

$$R(X_1, X_2) = b_0 + b_1 X_1 + b_2 X_2.$$

Применение соотношения (8.3) дает: $b_0 = 88{,}5$ МПа, $b_1 = 13$ МПа/ч, $b_2 = 7 \cdot 10^5$ МПа/кг. Модель имеет вид

$$R = 88{,}5 + 13X_1 + 7 \cdot 10^5 X_2.$$

Значение коэффициента при массе модификатора X_2 на *пятья порядков* превышает значение коэффициента при времени тепловой обработки X_1. Однако заключение о преимущественном влиянии массы модификатора на прочность будет ошибочным уже только по причине того, что коэффициенты вообще нельзя сравнивать – они имеют различные единицы измерения! Более того, даже при совпадении единиц измерения анализ влияния факторов на прочность

[1] Заметим, что в подобных технологических задачах выбирать в качестве входной переменной *массу* не принято; варьированию подвергаются массовые или объемные *доли* компонент. Однако суть примера полнее раскрывается именно в такой формулировке.

нельзя сделать на основании сравнений значений коэффициентов модели – дело в резком различии порядков действующих переменных.

Перейдем к нормализованным переменным. Основные уровни:

$$X_{1,0} = \frac{4+2}{2} = 3 \text{ ч}, \quad X_{2,0} = \frac{2\cdot 10^{-5} + 1\cdot 10^{-5}}{2} = 1{,}5\cdot 10^{-5} \text{ кг}.$$

Интервалы варьирования действующих переменных:

$$\Delta X_1 = \frac{4-2}{2} = 1 \text{ ч}, \quad \Delta X_2 = \frac{2\cdot 10^{-5} - 1\cdot 10^{-5}}{2} = 0{,}5\cdot 10^{-5} \text{ кг}.$$

План эксперимента в нормализованном факторном пространстве (иначе – план эксперимента в *кодовом* выражении) и соответствующие значения отклика приведены в табл. 9.2.

Таблица 9.2

x_1	x_2	R, МПа
−1	−1	121
−1	1	129
1	−1	148
1	1	154

ЭС-модель прочности по форме остается неизменной:

$$R(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 .$$

Оценки параметров: $\beta_0 = 138$ МПа, $\beta_1 = 13$ МПа, $\beta_2 = 3{,}5$ МПа. Значения коэффициентов модели указывают на доминирующее влияние *времени тепловой обработки*.

В приведенном примере ковариационная матрица исходного плана

$$\mathbf{D}_1 \approx \begin{pmatrix} 4{,}75 & -0{,}75 & -1{,}5\cdot 10^{5} \\ -0{,}75 & 0{,}25 & \sim 10^{-11} \\ -1{,}5\cdot 10^{5} & \sim 10^{-11} & 1\cdot 10^{10} \end{pmatrix}$$

свидетельствует, во-первых, о различной точности определения коэффициентов (оценки их дисперсий различаются на 11 порядков) и, во-вторых, о коррелированности коэффициентов друг с другом (отличие от нуля внедиагональных элементов ковариационной матрицы). Для той же модели в нормализованном факторном пространстве матрица ошибок равна

$$\mathbf{D}_2 = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{pmatrix}.$$

План эксперимента лишен двух указанных недостатков.

9.1. Использование средств MS Excel для построения квадратичной модели в нормализованном факторном пространстве

Пусть в процессе эксперимента варьируются два фактора. Известно, что отклик линейно зависит от первого фактора и квадратично – от второго. Исходя из этого выбрана модель

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 .$$

Для нахождения параметров модели выполнен эксперимент, план которого (в натуральных переменных) вместе с эмпирическими значениями отклика приведен в табл. 9.3.

Таблица 9.3

X_1	X_2	y
10	5	18
20	5	6
10	10	12
20	10	8
10	15	16
20	15	20

Требуется перейти к нормализованным переменным и найти коэффициенты ЭС-модели.

	A	B	C
1	10	5	18
2	20	5	6
3	10	10	12
4	20	10	8
5	10	15	16
6	20	15	20

Пусть значения действующих переменных помещены в первые два столбца рабочего листа, значения отклика помещены в третий столбец. Основные уровни и интервалы варьирования в данном примере равны:

$$X_{1,0} = \frac{20+10}{2} = 15; \quad X_{2,0} = 10;$$

$$\Delta X_1 = \frac{20-10}{2} = 5; \quad \Delta X_2 = 5.$$

Пусть они записаны в ячейки от `A8` до `A11`. Для нахождения мат-матрицы плана в кодовом

	A	B	C	D	E
1	10	5	18		
2	20	5	6		
3	10	10	12		
4	20	10	8		
5	10	15	16		
6	20	15	20		
7					
8	15	Основной уровень первого фактора			
9	10	Основной уровень второго фактора			
10	5	Интервал варьирования первого фактора			
11	5	Интервал варьирования второго фактора			
12					
13	Матрица плана в кодовом выражении				
14	=(A1-$A$8)/$A$10				
15	1	-1			
16	-1	0			
17	1	0			
18	-1	1			
19	1	1			

выражении поместим в две соседние ячейки какой-либо строки (в данном примере – строка 14) формулы

```
=(A1-$A$8)/$A$10
=(B1-$A$9)/$A$11
```

и переместим маркер автозаполнения на шесть строк ниже.

Базисные функции выбранной модели:

$$\varphi_1 = 1,\ \varphi_2 = x_1,\ \varphi_3 = x_2,$$

$$\varphi_4 = x_1 x_2,\ \varphi_5 = x_2^2.$$

Составим матрицу базисных функций. В пять соседних ячеек строки 22 поместим формулы

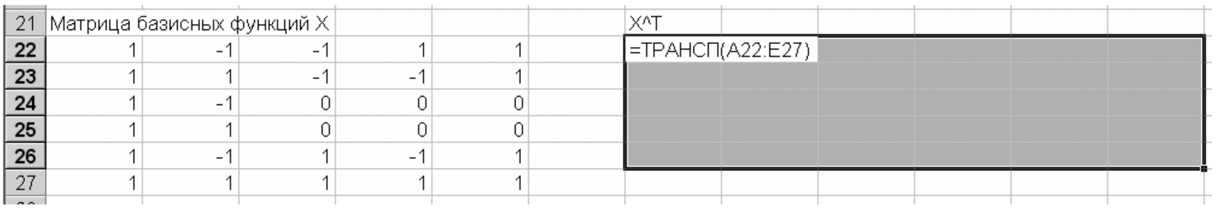
```
=1
=A14
=B14
=A14*B14
=B14*B14
```

После этого выделим диапазон A22:E22 и переместим маркер автозаполнения на шесть строк ниже.

Для использования соотношения (8.3) потребуется матрица $\mathbf{X}^T$, полученная транспонированием матрицы базисных функций. Выделим диапазон G22:L26 начиная с ячейки G22. Затем (при активном выделении) следует поместить в ячейку G22 формулу

```
=ТРАНСП(A22:E27)
```

и нажать Ctrl+Shift+Enter.



21	Матрица базисных функций X						X^T					
22	1	-1	-1	1	1		=ТРАНСП(A22:E27)					
23	1	1	-1	-1	1							
24	1	-1	0	0	0							
25	1	1	0	0	0							
26	1	-1	1	-1	1							
27	1	1	1	1	1							

Найдем матрицу моментов и ковариационную матрицы. В данном примере ЭС-модель содержит пять слагаемых, поэтому матрицы $\mathbf{X}^T\mathbf{X}$ и $\mathbf{D} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ имеют размеры 5x5. Можно выделить диапазон A30:E34, при активном выделении ввести в ячейку A30 формулу

```
=МУМНОЖ(G22:L26;A22:E27)
```

и нажать Ctrl+Shift+Enter.

21	Матрица базисных функций X				
22	1	-1	-1	1	1
23	1	1	-1	-1	1
24	1	-1	0	0	0
25	1	1	0	0	0
26	1	-1	1	-1	1
27	1	1	1	1	1
28					
29	Информационная матрица X^T*X				
30	=МУМНОЖ(G22:L26;A22:E27)				
31					
32					
33					
34					

Для нахождения ковариационной матрицы следует выделить диапазон `G30:K34`, при активном выделении ввести в ячейку `G30` формулу

`=МОБР(A30:E34)`

и нажать `Ctrl+Shift+Enter`.

29	Информационная матрица X^T*X						Ковариационная матрица				
30	6	0	0	0	4		=МОБР(A30:E34)		0	0	- 1/2
31	0	6	0	0	0		0	1/6	0	0	0
32	0	0	4	0	0		0	0	1/4	0	0
33	0	0	0	4	0		0	0	0	1/4	0
34	4	0	0	0	4		- 1/2	0	0	0	3/4

Заключительной операцией является нахождение вектора параметров модели. Можно выделить диапазон `G2:G6`, при активном выделении поместить в ячейку `G2` формулу

`=МУМНОЖ(МУМНОЖ(G30:K34;G22:L26);C1:C6)`

и нажать `Ctrl+Shift+Enter`.

	A	B	C	D	E	F	G	H	I	J	K	L
1	10	5	18				Вектор коэффициентов					
2	20	5	6				=МУМНОЖ(МУМНОЖ(G30:K34;G22:L26);C1:C6)					
3	10	10	12				-2					
4	20	10	8				3					
5	10	15	16				4					
6	20	15	20				5					
7												
8	15	Основной уровень первого фактора										
9	10	Основной уровень второго фактора										
10	5	Интервал варьирования первого фактора										
11	5	Интервал варьирования второго фактора										
12												
13	Матрица плана в кодовом выражении											
14	-1	-1										
15	1	-1										
16	-1	0										
17	1	0										
18	-1	1										
19	1	1										
20												
21	Матрица базисных функций X						X^T					
22	1	-1	-1	1	1		1	1	1	1	1	1
23	1	1	-1	-1	1		-1	1	-1	1	-1	1
24	1	-1	0	0	0		-1	-1	0	0	1	1
25	1	1	0	0	0		1	-1	0	0	-1	1
26	1	-1	1	-1	1		1	1	0	0	1	1
27	1	1	1	1	1							
28												
29	Информационная матрица X^T*X						Ковариационная матрица					
30	6	0	0	0	4		1/2	0	0	0	- 1/2	
31	0	6	0	0	0		0	1/6	0	0	0	
32	0	0	4	0	0		0	0	1/4	0	0	
33	0	0	0	4	0		0	0	0	1/4	0	
34	4	0	0	0	4		- 1/2	0	0	0	3/4	

Вектор параметров равен

$$\mathbf{B} = (10,-2,3,4,5)^T.$$

Искомая модель имеет вид

$$y = 10 - 2x_1 + 3x_2 + 4x_1x_2 + 5x_2^2.$$

Графическое представление предсказанного моделью значения отклика – поверхность отклика в трехмерном пространстве (x_1, x_2, y) или же линии равного отклика на плоскости (x_1, x_2) – рис. 9.2. К сожалению, MS Excel ограниченно пригоден для выполнения подобных построений; графическое представление результатов следует выполнять другими средствами.

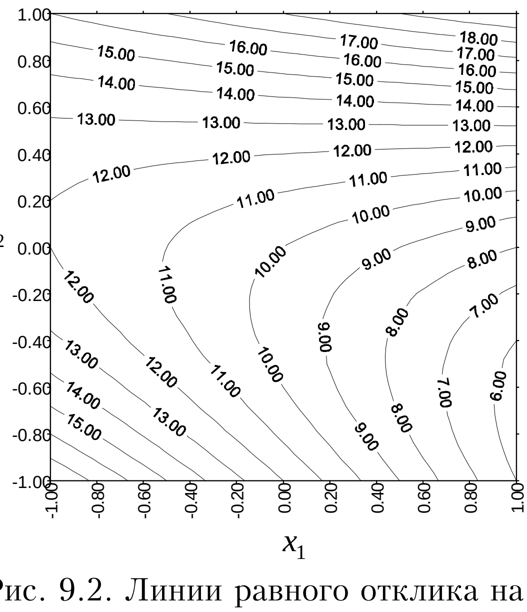


Рис. 9.2. Линии равного отклика на плоскости нормализованных переменных

10. Построение планов экспе-

римента

Активный эксперимент подразумевает определенную свободу выбора значений входных переменных. В рамках математической теории эксперимента выработаны рекомендации, позволяющие для данной экспериментально-статистической модели выбрать уровни варьируемых факторов так, чтобы обеспечить выполнение тех или иных свойств плана эксперимента – сделать эксперимент в определенном смысле *оптимальным*. Вновь отметим, что вопрос оптимальности плана существенно зависит от общего вида ЭС-модели и может решаться только *после* выбора последней.

Пусть ЭС-модель содержит L неизвестных параметров; тогда для их определения требуется, чтобы матрица плана содержала, как минимум, $N = L$ различных строк; подобные планы называют *насыщенными*. Если модель построена по насыщенному плану, то статистический анализ соответствия модели и экспериментальных данных без привлечения дополнительной информации (опытов в т.н. *контрольных точках*) выполнить невозможно.

Критерии оптимальности плана можно разделить на две группы: критерии, связанные с дисперсией оценок параметров, и критерии, связанные с дисперсией предсказанных значений отклика. Среди критериев первой группы важным является критерий *ортогональности*; среди критериев второй – критерий *ротатабельности*.

Характеристики плана эксперимента определяются входящей в соотношение (8.3) ковариационной матрицей – матричным аналогом дисперсии. Записанная в безразмерных нормализованных переменных матрица плана не зависит от содержательной стороны исследования. Поэтому выбор плана эксперимента – это задача построения такой матрицы плана, для которой соответствующая ковариационная матрица (не зависящая от результатов эксперимента, но зависящая от выбранной ЭС-модели!)

$$\mathbf{D} = \mathbf{M}^{-1} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1L} \\ c_{21} & c_{22} & \dots & c_{2L} \\ \dots & & & \\ c_{L1} & c_{L2} & \dots & c_{LL} \end{pmatrix}$$

обладала бы определенными свойствами.

Ковариационная матрица определяет не только численные значения параметров модели, но и точность оценки этих параметров.

Диагональные элементы матрицы определяют дисперсии оценок параметров

$$s_{b_i}^2 = s_e^2 c_{ii}\,,$$

где s_e^2 – дисперсия эксперимента[1] (*дисперси воспроизводимости*). Если все диагональные элементы равны между собой, то точность оценки всех параметров будет одинакова.

Внедиагональные элементы определяют ковариацию (т.е. взаимное влияние) параметров. Если все внедиагональные элементы равны нулю, то параметры модели определяются независимо друг от друга; соответствующие планы эксперимента называют *ортогональными*.

Условием ортогональности плана является ортогональность столбцов его матрицы базисных функций – скалярное произведение двух любых различных столбцов этой матрицы должно быть нулевым:

$$\sum_{u=1}^{N} \varphi_i(\mathbf{x}_u)\varphi_j(\mathbf{x}_u) = 0\,,\ \ i, j = \overline{1, L}\,,\ \ i \neq j\,.$$

Примером ортогонального плана для линейной k-факторной модели является план *полного факторного эксперимента* (ПФЭ) 2^k. Экспериментальные точки ПФЭ 2^k расположены в вершинах k-мерного гиперкуба с центром в начале координат и длиной ребра, равной 2. Например, для двух факторов матрица плана 2^2 имеет вид

$$\begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{pmatrix}^T. \tag{10.1}$$

И матрица моментов, и ковариационная матрица этого плана для линейной модели

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

являются диагональными

$$\left(\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{pmatrix}^T \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \right)^{-1} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}^{-1} = \frac{1}{4}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

и свидетельствуют об ортогональности плана.

---

[1] См. соотношение (11.1).

Критерием, на основании которого может быть сделан вывод о количестве информации, содержащейся в предсказанном значении отклика, является связанная с ковариационной матрицей **D** *информационна - ункци* плана эксперимента:

$$I_x = \frac{1}{d} = \left(\mathbf{x}_p^T \mathbf{D} \mathbf{x}_p\right)^{-1}, \tag{10.2}$$

где $\mathbf{x}_p$ – вектор-столбец, образованный значениями базисных функций в соответствующей точке факторного пространства:

$$\mathbf{x}_p = \left(\varphi_1(\mathbf{x}), \varphi_1(\mathbf{x}), \ldots, \varphi_L(\mathbf{x})\right)^T.$$

Выражение

$$d = \mathbf{x}_p^T \mathbf{D} \mathbf{x}_p \tag{10.3}$$

называют *нормализованной неопределенностью*.

Если значение информационной функции (10.2) зависит только от расстояния между точкой **x** и центром исследуемой факторной области (для планов в безразмерных нормализованных переменных это, как правило, начало координат), то план эксперимента называют *ротатабельным*.

Например, для рассмотренного выше плана полного факторного эксперимента 2^2 имеем:

$$\mathbf{D} = \frac{1}{4}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$d = \frac{1}{4}\begin{pmatrix} 1 & x_1 & x_2 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1 & x_1 & x_2 \end{pmatrix}\begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} = \frac{1}{4}\left(1 + x_1^2 + x_2^2\right),$$

$$I_x = \frac{1}{d} = \frac{4}{1 + x_1^2 + x_2^2} = \frac{4}{1 + r^2},$$

где $r = \sqrt{x_1^2 + x_2^2}$ – расстояние от начала координат то точки $\left(x_1, x_2\right)$. Информационная функция радиально-симметрична, и план 2^2 для линейной модели является ротатабельным.

Нетрудно проверить, что свойство ротатабельности плана 2^2 сохраняется при его повороте вокруг начала координат. Так, план

$$\begin{pmatrix} 0 & -\sqrt{2} & 0 & \sqrt{2} \\ \sqrt{2} & 0 & -\sqrt{2} & 0 \end{pmatrix}^T, \tag{10.4}$$

полученный из 2^2 поворотом на угол $\pi/4$, имеет ковариационную матрицу

$$\left(\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & -\sqrt{2} & 0 & \sqrt{2} \\ \sqrt{2} & 0 & -\sqrt{2} & 0 \end{pmatrix}^T \begin{pmatrix} 1 & 0 & \sqrt{2} \\ 1 & -\sqrt{2} & 0 \\ 1 & 0 & -\sqrt{2} \\ 1 & \sqrt{2} & 0 \end{pmatrix}\right)^{-1} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}^{-1} = \frac{1}{4}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

.

Совпадение этой матрицы с ковариационной матрицей плана 2^2 свидетельствует о том, что свойства этих планов эксперимента полностью одинаковы. Тем не менее, простые соображения о требуемом диапазоне варьирования факторов свидетельствуют о преимуществе плана полного факторного эксперимента 2^2: для плана (10.4) размах варьирования увеличен в $\sqrt{2}$ раз по сравнению с исходным планом (10.1).

Очевидно, что в случае вырожденности матрицы моментов ковариационную матрицу (и оценки параметров) найти невозможно. Уменьшение абсолютной величины определителя матрицы моментов[1] сопровождается возрастанием абсолютных величин элементов ковариационной матрицы и увеличением ошибок определения коэффициентов модели[2].

---

[1] Точнее, ухудшение ее *обусловленности*.

[2] Существуют методы, позволяющие в некотором смысле получить «решение» системы линейных уравнений даже тогда, когда система несовместна. В частности, методы, основанные на *сингул рном разложении* матрицы коэффициентов, позволяют получить «решение» именно в смысле метода наименьших квадратов. Однако применение этих (весьма сложных!) методов вряд ли оправдано – близость матрицы моментов к вырожденности свидетельствует только о некорректном выборе плана и/или вида ЭС-модели.

Вырожденность матрицы моментов имеет место тогда, когда при выбранном плане эксперимента «базисные» функции ЭС-модели на самом деле таковыми не являются (столбцы матрицы базисных функций линейно зависимы). Этот случай можно проиллюстрировать на примере пятиточечного плана эксперимента с матрицей

$$\begin{pmatrix} 0 & 1 & -1 & 1 & -1 \\ 0 & 1 & 1 & -1 & -1 \end{pmatrix}^T;$$

план изображен на рис. 10.1.

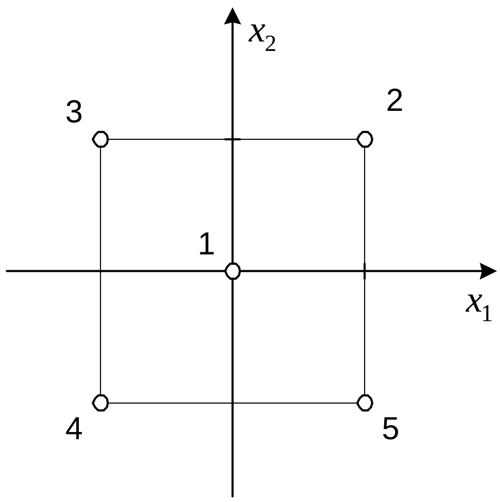


Рис. 10.1. План, непригодный для построения квадратичной модели

На первый взгляд задача построения квадратичной ЭС-модели

$$y = \beta_0 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

по результатам эксперимента в соответствии с планом на рис. 10.1 должна решаться однозначно. Действительно, данная задача состоит в минимизации суммы квадратов отклонений экспериментальных точек (x_{1u}, x_{2u}, y_u) от эллиптического параболоида. Три параметра определяются по результатам экспериментов, выполненных в пяти различных точках. Однако для выбранных плана и модели матрица базисных функций

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

содержит два совпадающих столбца, и матрица моментов

$$\mathbf{M} = \mathbf{X}^T \mathbf{X} = \begin{pmatrix} 5 & 4 & 4 \\ 4 & 4 & 4 \\ 4 & 4 & 4 \end{pmatrix}$$

оказывается вырожденной!

Дальнейший анализ содержания задачи позволяет выявить ее неопределенность. Равенство двух последних столбцов матрицы базисных функций приводит к тому, что значение модели в каждой из экспериментальных точек определяется уровнем только одной из переменных (говорят, что в данном случае имеет место *смешанна оценка*):

$$y = \beta_0 + \beta_{11} x_1^2 + \beta_{22} x_2^2 = \beta_0 + (\beta_{11} + \beta_{22}) x_1^2 .$$

Поэтому если для двух моделей

$$y_1 = a + b_1 x_1^2 + c_1 x_2^2 , \;\; y_2 = a + b_2 x_1^2 + c_2 x_2^2$$

выполнено

$$b_1 + c_1 = b_2 + c_2 ,$$

то значения этих моделей в каждой из пяти точек плана на рис. 10.1 будут одинаковы.

Заметим, что уже простой разворот плана на угол $\pi/4$ вокруг начала координат (рис. 10.2) решает проблему вырожденности матрицы моментов. Столбцы матрицы базисных функций

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

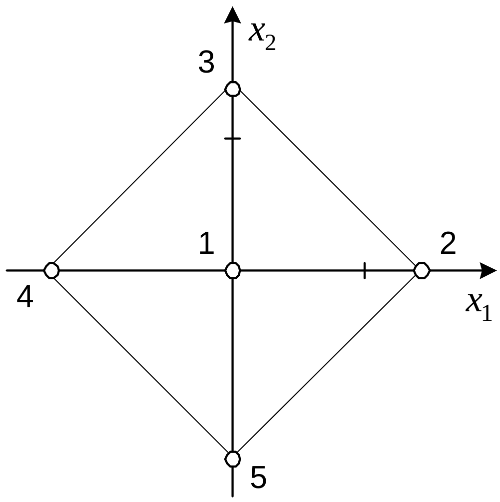


Рис. 10.2

полученного плана линейно независимы, поэтому матрица моментов

$$\mathbf{M} = \begin{pmatrix} 5 & 4 & 4 \\ 4 & 8 & 0 \\ 4 & 0 & 8 \end{pmatrix}$$

вырожденной не является. Ковариационная матрица существует и равна

$$\mathbf{D} = \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 3/8 & 1/4 \\ -1/2 & 1/4 & 3/8 \end{pmatrix} .$$

Полученный план не является ортогональным. Очевидно также, что он не может быть ротатабельным. Действительно,

$$d=\begin{pmatrix}1\\x_1^2\\x_2^2\end{pmatrix}^T\begin{pmatrix}1&-1/2&-1/2\\-1/2&3/8&1/4\\-1/2&1/4&3/8\end{pmatrix}\begin{pmatrix}1\\x_1^2\\x_2^2\end{pmatrix}=\begin{pmatrix}1\\x_1^2\\x_2^2\end{pmatrix}^T\begin{pmatrix}1-\dfrac{x_1^2}{2}-\dfrac{x_2^2}{2}\\-\dfrac{1}{2}+\dfrac{3}{8}x_1^2+\dfrac{1}{4}x_2^2\\-\dfrac{1}{2}+\dfrac{1}{4}x_1^2+\dfrac{3}{8}x_2^2\end{pmatrix}=$$

$$=1-\frac{x_1^2}{2}-\frac{x_2^2}{2}-\frac{x_1^2}{2}+\frac{3}{8}x_1^4+\frac{1}{4}x_1^2x_2^2-\frac{x_2^2}{2}+\frac{1}{4}x_1^2x_2^2+\frac{3}{8}x_2^4=$$

$$=1-\left(x_1^2+x_1^2\right)+\frac{1}{2}x_1^2x_2^2+\frac{3}{8}\left(x_1^4+x_2^4\right).$$

Ни нормализованная неопределенность d, ни информационная функция (рис. 10.3) не являются радиально-симметричными.

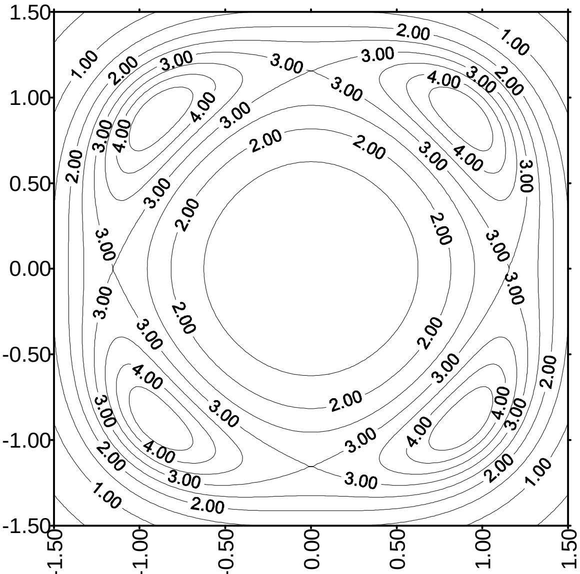


Рис. 10.3. Информационная функция плана (см. рис. 10.2)

11. Анализ моделей, линейных по параметрам

Вычислительная процедура построения линейной по параметрам ЭС-модели сводится к использованию соотношения (8.3). Важно, что это соотношение позволяет лишь найти коэффициенты модели, но *не решает вопроса соответствияя построеннойя моделииябъектаяисследования*.

Как уже было отмечено, в основе метода наименьших квадратов лежат три предположения: предположение о нормальном распределении ошибок, о независимости и равной точности измерений. Если хотя бы одно из них нарушено, то применение метода наименьших квадратов недопустимо: полученная этим методом ЭС-модель будет плохим описанием объекта (предсказанные моделью значения будут далеки от истинных).

Поэтому проверка предположений должна выполняться *до* построения модели; она становится возможной только тогда, когда в каждом u-м из $u=\overline{1,N}$ экспериментов измерение значения отклика повторяется $m_u>1$ раз. Эти m_u измерений, соответствующие *одной* экспериментальной точке, называют *параллельными*.

Предположение о независимости измерений проверить непосредственно *в одном эксперименте* невозможно. С целью снижения возможной взаимной зависимости обычно выполняют рандомизацию измерений (проводят измерения в случайном порядке).

Проверка гипотезы о нормальном распределении ошибок требует получения выборки большого объема (по крайней мере – около 100), поэтому на практике ограничиваются только проверкой гипотезы о равной точности измерений.

Пусть в u-й точке выполнено m_u параллельных измерений. Тогда в каждом эксперименте выборочные средние и выборочные дисперсии находятся как

$$\bar{y}_u = \frac{1}{m_u}\sum_{i=1}^{m_u} y_{ui}\,,\ \ u = \overline{1,N}\,,$$

$$s_u^2 = \frac{1}{m_u - 1}\sum_{i=1}^{m_u}\left(y_{ui} - \bar{y}_u\right)^2 = \frac{1}{f_u}\sum_{i=1}^{m_u}\left(y_{ui} - \bar{y}_u\right)^2\,,\ \ u = \overline{1,N}\,,$$

где $f_u = m_u - 1$ – число *степеней свободы* выборочной дисперсии (число параллельных испытаний, уменьшенное на число найденных по выборке оценок – при вычислении выборочной дисперсии уже найдено выборочное среднее). Первый индекс u в обозначении отклика y_{ui} является номером эксперимента, второй индекс i – номером параллельного испытания в этом эксперименте.

В предположении о равной точности найденные оценки выборочных дисперсий позволяют вычислить *дисперсию воспроизводимости* (*дисперсию эксперимента*):

$$s_e^2 = \frac{f_1 s_1^2 + f_2 s_2^2 + \ldots + f_N s_N^2}{f_1 + f_2 + \ldots + f_N} = \frac{1}{f_e}\sum_{u=1}^{N} f_u s_u^2 = \frac{1}{f_e}\sum_{u=1}^{N}\sum_{i=1}^{m_u}\left(y_{ui} - \bar{y}_u\right)^2\,, \qquad (11.1)$$

где

$$f_e = \sum_{u=1}^{N} f_u = \sum_{u=1}^{N}\left(m_u - 1\right) = \sum_{u=1}^{N} m_u - N$$

– число степеней свободы дисперсии воспроизводимости (полное число измерений, включая параллельные, за вычетом числа экспериментов в различных точках).

Если в каждом эксперименте число параллельных измерений одинаково

$$m_1 = m_2 = \ldots = m_N = M\,,$$

то соотношение (11.1) упрощается:

$$s_e^2 = \frac{1}{MN - N}\sum_{u=1}^{N}(M-1)s_u^2 = \frac{1}{N}\sum_{u=1}^{N}s_u^2\ ; \qquad (11.2)$$

дисперсия воспроизводимости вычисляется как *среднее арифметическое* всех выборочных дисперсий.

Проверить гипотезу о равенстве генеральных дисперсий для всех $u = \overline{1,N}$ экспериментов при $m_u > 3$ можно по *критерию Бартлета*. Вычисляются величины

$$B = 2{,}303\left(f_e \lg s_e^2 - \sum_{u=1}^{N}(m_u - 1)\lg s_u^2\right);$$

$$C = 1 + \frac{1}{3(N-1)}\left[\left(\sum_{u=1}^{N}\frac{1}{m_u - 1}\right) - \frac{1}{f_e}\right].$$

Затем отыскивается значение статистики B/C. Можно приближенно считать, что данная статистика подчинена χ^2-распределению с $N-1$ степенями свободы.

[1]При совпадающем числе параллельных испытаний наиболее удобным способом проверки гипотезы о равенстве генеральных дисперсий оказывается *G-критерий* (*критерийяКохрена*). Для его использовании вычисляется статистика

$$G = \max\left\{s_u^2\right\}\Big/\sum_{u=1}^{N}s_u^2\ , \qquad (11.3)$$

равная отношению максимальной из выборочных дисперсий к сумме всех выборочных дисперсий. Статистика (11.3) подчинена *G-распределению* (*распределениюя Кохрена*) со степенями свободы $f_1 = M - 1$ и $f_2 = N$.

Эмпирическое значение (11.3) G-статистики позволяет найти вероятность критического события, состоящего в том, что в условиях равной точности измерений неизвестное истинное отношение

$$\max\left\{\sigma_u^2\right\}\Big/\sum_{u=1}^{N}\sigma_u^2$$

окажется столь же большим, как в эксперименте. Если вероятность критического события оказывается меньше заданного уровня значимости (как правило, $\alpha = 0{,}05$), то гипотеза о равенстве дисперсий в

[1] UPD (2012, kkatarn): несколько следующих абзацев являются бредом) «Распределения» Кохрена (Cochran) не существует, есть только критерий)

каждом из N экспериментов отвергается и построение ЭС-модели оказывается невозможным.

Обычно в распоряжении исследователя имеются только таблицы квантилей G-распределения. В этом случае найденное эмпирическое значение G-статистики следует сравнить с квантилью G_{α, f_1, f_2} распределения Кохрена для выбранного уровня значимости α и числа степеней свободы $f_1 = M - 1$ и $f_2 = N$. При выполнении неравенства

$$G < G_{\alpha, f_1, f_2}$$

гипотеза о равной точности измерений не отвергается.

После проверки однородности дисперсий параллельных опытов и отыскания коэффициентов ЭС-модели для каждого из найденных коэффициентов необходимо проверить гипотезу о равенстве истинного значения коэффициента нулю. Если в условиях эксперимента отвергать данную гипотезу нет оснований, то говорят, что коэффициент *статистически незначим*.

Для проверки значимости коэффициента β_j, $j = \overline{1, L}$ находят значение статистики

$$t_j = \frac{\beta_j}{\sqrt{s_e^2 c_{jj}}}, \tag{11.4}$$

где s_e^2 – дисперсия воспроизводимости, c_{jj} – диагональный элемент ковариационной матрицы. Статистика (11.4) подчинена распределению Стьюдента с $f = f_e = N(M-1)$ степенями свободы (в случае неортогональных планов это выполнено лишь приближенно).

Гипотеза о равенстве нулю неизвестного истинного значения j-го параметра должна быть отвергнута в пользу двусторонней альтернативы, если вероятность

$$p_j = \int_{-\infty}^{-|t_j|} f(x)dx + \int_{|t_j|}^{\infty} f(x)dx = 1 - 2\int_0^{|t_j|} f(x)dx \tag{11.5}$$

критического события, состоящего в том, что при указанной гипотезе будет получено значение β_j, большее или равное найденного в эксперименте, оказывается меньше заданного уровня значимости α (в соотношении (11.5) подынтегральная функция является плотностью распределения Стьюдента).

Если имеются таблицы квантилей $t_{N(M-1),\alpha}$ распределения Стьюдента для $N(M-1)$ степеней свободы и выбранного уровня значимости α, то с квантилью следует сравнить абсолютную величину статистики (11.4). При выполнении неравенства

$$\left|t_j\right| \geq t_{N(M-1),\alpha}$$

гипотеза о статистической незначимости параметра отвергается.

Все незначимые коэффициенты ЭС-модели обнуляются; это, очевидно, соответствует отбрасыванию некоторых слагаемых ЭС-модели. Если ковариационная матрица не является диагональной (план не был ортогональным), то оставшиеся коэффициенты необходимо пересчитать заново.

Последнее определяет *итерационный процесс регрессионного анализа*: наличие статистически незначимых оценок параметров ЭС-модели требует изменения ее вида, повторного отыскания коэффициентов и последующей проверки статистической значимости каждого из них.

Заключительным шагом анализа является проверка *адекватности* полученной ЭС-модели результатам эксперимента. Для ее выполнения вычисляется *остаточнаядисперсия*, или *дисперсияадекватности* –величина, пропорциональная сумме квадратов разностей между предсказанными моделью и эмпирическими значениями отклика. Если в каждой точке факторного пространства выполняется M параллельных измерений, то дисперсия адекватности равна

$$s_{ad}^2 = \frac{M}{N-L}\sum_{u=1}^{N}\left(y_u - f\left(\mathbf{x}_u\right)\right)^2, \tag{11.6}$$

где N – число различных экспериментов (число точек плана эксперимента), L – число искомых параметров модели.

Затем вычисляется значение статистики

$$F = \frac{s_{ad}^2}{s_e^2}, \tag{11.7}$$

где s_e^2 – дисперсия воспроизводимости. Статистика (11.7) подчинена распределению Фишера с $f_{ad} = N - L$ и $f_e = N(M-1)$ степенями свободы. Гипотеза адекватности модели эксперименту отвергается, если вероятность

$$p = \int_F^\infty f(x)dx = 1 - \int_0^F f(x)dx = 1 - G(F)$$

критического события, состоящего в том, что при адекватной модели значение F будет столь же большим, как и в эксперименте, окажется меньше выбранного уровня значимости (в последнем соотношении $f(x)$ и $G(x)$ – плотность и функция распределения Фишера, соответственно).

Если имеются таблицы квантилей $F_{N-L,N(M-1),\alpha}$ распределения Фишера для $N-L$, $N(M-1)$ степеней свободы и выбранного уровня значимости α, то статистика (11.7) сравнивается с квантилью. При выполнении неравенства

$$F < F_{N-L,N(M-1),\alpha}$$

гипотеза адекватности модели не отвергается.

11.1. Построение и анализ линейной двухфакторной модели

Пусть требуется построить линейную двухфакторную экспериментально-статистическую модель

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 .$$

Для нахождения ее параметров поставлен полный факторный эксперимент 2^2 с числом опытов $N=4$. В каждой точке факторного пространства выполнено $M=3$ параллельных испытания. План эксперимента (в нормализованных переменных), вместе с эмпирическими значениями отклика, приведен в табл. 11.6.

Таблица 11.1

Уровни входных переменных		Значения отклика в параллельных испытаниях		
x_1	x_2	y_1	y_2	y_3
−1	−1	−2,04691	−2,00964	−2,02786
1	−1	0,019879	0,070023	−0,06276
−1	1	2,081853	1,918242	2,098057
1	1	4,057741	3,91337	4,046098

Требуется:

1. Проверить статистическую гипотезу о равной точности измерений в каждой серии параллельных испытаний.

2. В том случае, если измерения равноточны, найти оценки параметров модели.

3. Выяснить, какие из параметров являются статистически значимыми, а какие – нет. Обнулить незначимые коэффициенты (пересчет значимых при этом можно не производить, так как выбранный план эксперимента является ортогональным).

4. Проверить статистическую гипотезу об адекватности построенной модели экспериментальным данным.

Запишем матрицу базисных функций в ячейки `A3:C6`, эмпирические значения отклика – в ячейки `D3:F6`.

	A	B	C	D	E	F
1					Y	
2		X		1	2	3
3	1	1	-1	-2,04691	-2,00964	-2,02786
4	1	-1	-1	0,019879	0,070023	-0,06276
5	1	1	1	2,081853	1,918242	2,098057
6	1	-1	1	4,057741	3,91337	4,046098

Для каждой серии параллельных испытаний найдем оценки математического ожидания и дисперсии. В ячейки `G3` и `H3` введем формулы

```
=СУММ(D3:F3)/3
=((D3-G3)^2+(E3-G3)^2+(F3-G3)^2)/2
```

выделим диапазон `G3:H3` и переместим маркер автозаполнения до шестой строки.

	B	C	D	E	F	G	H	I
1				Y		Yav	Sy^2	
2	X		1	2	3			
3	1	-1	-2,0469140	-2,0096415	-2,0278639	=СУММ(D3:F3)/3	=((D3-G3)^2+(E3-G3)^2+(F3-G3)^2)/2	
4	-1	-1	0,01987860	0,07002290	-0,0627594			
5	1	1	2,08185276	1,91824211	2,0980567			
6	-1	1	4,05774066	3,91336970	4,0460981			

Так как число испытаний в каждой серии параллельных опытов одинаково, то для проверки однородности дисперсий целесообразно применить G-критерий Кохрена.

Найдем дисперсию воспроизводимости и максимальную из дисперсий параллельных опытов. В ячейку `H7` введем

```
=СУММ(H3:H6)/4
```

В ячейку `H8` введем

```
=МАКС(H3;H4;H5;H6)
```

	A	B	C	D	E	F	G	H	I
1					Y		Yav	Sy^2	
2		X		1	2	3			
3	1	1	-1	-2,04691	-2,00964	-2,02786	-2,02814	0,000347	
4	1	-1	-1	0,019879	0,070023	-0,06276	0,009047	0,004496	
5	1	1	1	2,081853	1,918242	2,098057	2,032717	0,009894	
6	1	-1	1	4,057741	3,91337	4,046098	4,005736	0,006433	
7					**Дисп. воспроизводимости**			0,005292	
8					**Макс. из дисперсий опытов**			=МАКС(H3;H4;H5;H6)	
9							**G**	0,467367	
10							**G_005**	0,768	
11					**Измерения равноточны (alpha=0,05)**				

Расчетное значение G-критерия

$$G \approx \frac{0{,}009894}{0{,}02117} \approx 0{,}4674$$

найдем, поместив в ячейку H9 формулу

```
=H8/(H7*4)
```

К сожалению, среди функций рабочего листа MS Excel нет функции, возвращающей значение G-распределения. Поэтому квантиль этого распределения для уровня значимости 0,05 и числа степеней свободы $f_1 = M - 1 = 3 - 1 = 2$, $f_2 = N = 4$ найдем по таблице; он равен

$$G_{0{,}05;2;4} = 0{,}768\,.$$

Поместим значение $G_{0{,}05;2;4} = 0{,}768$ в ячейку H10.

Так как расчетное значение G-критерия меньше 0,768, то экспериментальные данные не дают основания отвергать гипотезу о равной точности измерений. Этот вывод лучше зафиксировать непосредственно на рабочем листе; достаточно в ячейку H11 ввести

```
=ЕСЛИ($H$9<$H$10;"Измерения равноточны";"Измерения неравноточны")
```

	A	B	C	D	E	F	G	H
1					Y		Yav	Sy^2
2		X		1	2	3		
3	1	1	-1	-2,04691	-2,00964	-2,02786	-2,02814	0,000347
4	1	-1	-1	0,019879	0,070023	-0,06276	0,009047	0,004496
5	1	1	1	2,081853	1,918242	2,098057	2,032717	0,009894
6	1	-1	1	4,057741	3,91337	4,046098	4,005736	0,006433
7							Дисп. воспроизводимости	0,005292
8							Макс. из дисперсий опытов	0,009894
9							G	0,467367
10							G_005	0,768
11	=ЕСЛИ($H$9<$H$10;"Измерения равноточны";"Измерения неравноточны")							

Параметры модели можно найти, воспользовавшись общей формулой $\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ (использование этого соотношения здесь избыточно, так как и матрица моментов, и ковариационная матрицы плана 2^2 являются диагональными). Столбец $\mathbf{Y}$ образован средними значениями отклика в каждой серии параллельных испытаний.

Выделим диапазон A8:D10, при активном выделении введем в ячейку A8 формулу

```
=ТРАНСП(A3:C6)
```

завершая ввод нажатием на Ctrl+Shift+Enter. Затем найдем матрицу моментов $\mathbf{X}^T\mathbf{X}$ (ячейки A12:C14), ковариационную матрицу $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ (ячейки A16:C18), матрицу $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ (ячейки

`A20:D22`) и вектор коэффициентов $\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ (ячейки `E16:E18`)

	A	B	C	D
1				
2		X		1
3	1	1	-1	-2,04691
4	1	-1	-1	0,019879
5	1	1	1	2,081853
6	1	-1	1	4,057741
7			X^T	
8	1	1	1	1
9	1	-1	1	-1
10	-1	-1	1	1
11		M=X^T*X		
12	=МУМНОЖ(A8:D10;A3:C6)			
13	0	4	0	
14	0	0	4	

	A	B	C
1			
2		X	
3	1	1	-1
4	1	-1	-1
5	1	1	1
6	1	-1	1
7		X^T	
8	1	1	1
9	1	-1	1
10	-1	-1	1
11		M=X^T*X	
12	4	0	0
13	0	4	0
14	0	0	4
15		D=M^-1	
16	=МОБР(A12:C14)		0
17	0	0,25	0
18	0	0	0,25

	A	B	C	D	E	F	G
1					Y		Yav
2		X		1	2	3	
3	1	1	-1	-2,04691	-2,00964	-2,02786	-2,02814
4	1	-1	-1	0,019879	0,070023	-0,06276	0,009047
5	1	1	1	2,081853	1,918242	2,098057	2,032717
6	1	-1	1	4,057741	3,91337	4,046098	4,005736
7		X^T			Дисп. воспроизводимости		
8	1	1	1	1	Макс. из дисперсий опытов		
9	1	-1	1	-1			G
10	-1	-1	1	1			G_005
11		M=X^T*X			Измерения равноточны (al		
12	4	0	0				
13	0	4	0				
14	0	0	4				
15		D=M^-1			B=L*Y		
16	0,25	0	0		=МУМНОЖ(A20:D22;G3:G6)		
17	0	0,25	0		-1,00255		
18	0	0	0,25		2,014386		
19		L=D*X^T					
20	0,25	0,25	0,25	0,25			
21	0,25	-0,25	0,25	-0,25			
22	-0,25	-0,25	0,25	0,25			

При проверке статистической значимости параметров и адекватности модели выберем уровень значимости равным 0,05. Поместим это значение в ячейку `I2`.

Дисперсии коэффициентов модели равны

$$s_{b_i}^2 = s_e^2 c_{ii},$$

где s_e^2 – дисперсия воспроизводимости, c_{ii} – диагональные элементы ковариационной матрицы. Для нахождения дисперсий коэффициентов поместим в ячейки `F16`, `F17` и `F18` формулы

```
=$A$16*$H$7
=$B$17*$H$7
=$C$18*$H$7
```

Вычислим эмпирические значения t-критерия для каждого из коэффициентов. В ячейку `G16` введем

```
=ABS(E16)/F16^0,5
```

и переместим маркер автозаполнения до ячейки `G18`.

	A	B	C	D	E	F	G
15		D=M^-1			B=L*Y	Sb^2	tb
16	0,25	0	0		1,00484	0,001323	=ABS(E16)/F16^0,5
17	0	0,25	0		-1,00255	0,001323	27,56188
18	0	0	0,25		2,014386	0,001323	55,37897

Найдем вероятности критических событий – вероятности того, что при гипотезах $\beta_{j,ист} = 0$ значения параметров модели будут столь же велики, как в условиях поставленного эксперимента. В ячейку `H16` введем

```
=СТЬЮДРАСП(G16;8;2)
```

и переместим маркер автозаполнения до ячейки `H18`. В данном примере вероятности критических событий оказались $3 \cdot 10^{-9}$,

$3 \cdot 10^{-9}$ и 10^{-11}, что на несколько порядков меньше выбранного уровня значимости. Поэтому все три гипотезы о статистической незначимости соответствующих параметров отвергаются. Как и в случае проверки однородности дисперсий, данный результат можно фиксировать непосредственно на рабочем листе, используя логическую функцию `ЕСЛИ`. Эту же функцию можно использовать для обнуления статистически незначимых коэффициентов. Поместим в ячейки `I16` и `K16` формулы

```
=ЕСЛИ(H16<$I$2;"Параметр значим";"Параметр незначим")
=ЕСЛИ(H16<$I$2;E16;0)
```

и, после выделения диапазона `I16:K16`, переместим маркер автозаполнения до строки 18.

	A	B	C	D	E	F	G	H	I	J	K	L
1					Y		Yav	Sy^2	alpha			
2		X		1	2	3			0,05			
3	1	1	-1	-2,04691	-2,00964	-2,02786	-2,02814	0,000347				
4	1	-1	-1	0,019879	0,070023	-0,06276	0,009047	0,004496				
5	1	1	1	2,081853	1,918242	2,098057	2,032717	0,009894				
6	1	-1	1	4,057741	3,91337	4,046098	4,005736	0,006433				
7		X^T			Дисп. воспроизводимости			0,005292				
8	1	1	1	1	Макс. из дисперсий опытов			0,009894				
9	1	-1	1	-1			G	0,467367				
10	-1	-1	1	1			G_005	0,768				
11		M=X^T*X			Измерения равноточны (alpha=0,05)							
12	4	0	0									
13	0	4	0									
14	0	0	4									
15		D=M^-1			B=L*Y	Sb^2	tb	prob	Значимость параметров			
16	0,25	0	0		1,00484	0,001323	27,6248	3,18E-09	Параметр значим		=ЕСЛИ(H16<$I$2;E16;0)	
17	0	0,25	0		-1,00255	0,001323	27,56188	3,24E-09	Параметр значим		-1,00255	
18	0	0	0,25		2,014386	0,001323	55,37897	1,25E-11	Параметр значим		2,014386	

Последний шаг – проверка адекватности полученной ЭС-модели. Найдем предсказанные моделью значения отклика. В ячейку `L16` введем

```
=$K$16+$K$17*B3+$K$18*C3
```

и переместим маркер автозаполнения до строки 19.

15		D=M^-1			B=L*Y	Sb^2	tb	prob	Значимость параметров			Yf
16	0,25	0	0		1,00484	0,001323	27,6248	3,18E-09	Параметр значим		1,00484	=$K$16+$K$17*B3+$K$18*C3
17	0	0,25	0		-1,00255	0,001323	27,56188	3,24E-09	Параметр значим		-1,00255	-0,00699
18	0	0	0,25		2,014386	0,001323	55,37897	1,25E-11	Параметр значим		2,014386	2,016675
19		L=D*X^T										4,021778

Найдем дисперсию адекватности. В данном примере число параллельных испытаний $M = 3$, число опытов $N = 4$, число значимых коэффициентов модели $L = 3$. Поэтому дисперсия адекватности

$$s_{ad}^2 = \frac{M}{N-L} \sum_{u=1}^{N} \left(y_u - f\left(\mathbf{x}_u\right) \right)^2 = 3 \sum_{u=1}^{4} \left(y_u - f\left(\mathbf{x}_u\right) \right)^2 .$$

Поместим в ячейку `M16` формулу

```
=(L16-G3)^2
```

и переместим маркер автозаполнения до строки 19. Для нахождения эмпирического значения F-критерия поместим в ячейки `M20` и `M21` формулы

```
=3*СУММ(M16:M19)
=$M$20/$H$7
```

Найдем вероятность того, что отношение σ_{ad}^2/σ_e^2 (здесь σ_{ad}^2 и σ_e^2 – неизвестные истинные значения дисперсии адекватности и дисперсии воспроизводимости) окажется столь же большим, как в условиях эксперимента. Поместим в ячейку `M22` формулу

```
=FРАСП(M21;1;8)
```

и переместим маркер автозаполнения до строки 19. Для нахождения эмпирического значения F-критерия поместим в ячейки `M20` и `M21` формулы

```
=3*СУММ(M16:M19)
=$M$20/$H$7
```

Найдем вероятность того, что отношение σ_{ad}^2/σ_e^2 (здесь σ_{ad}^2 и σ_e^2 – неизвестные истинные значения дисперсии адекватности и дисперсии воспроизводимости) окажется столь же большим, как в условиях эксперимента. Поместим в ячейку `M22` формулу

```
=FРАСП(M21;1;8)
```


	A	B	C	D	E	F	G	H	I	J	K	L	M
1					Y		Yav	Sy^2	alpha				
2		X		1	2	3			0,05				
3	1	1	-1	-2,04691	-2,00964	-2,02786	-2,02814	0,000347					
4	1	-1	-1	0,019879	0,070023	-0,06276	0,009047	0,004496					
5	1	1	1	2,081853	1,918242	2,098057	2,032717	0,009894					
6	1	-1	1	4,057741	3,91337	4,046098	4,005736	0,006433					
7		X^T			Дисп. воспроизводимости			0,005292					
8	1	1	1	1	Макс. из дисперсий опытов			0,009894					
9	1	-1	1	-1			G	0,467367					
10	-1	-1	1	1			G_005	0,768					
11		M=X^T*X			Измерения равноточны (alpha=0,05)								
12	4	0	0										
13	0	4	0										
14	0	0	4										
15		D=M^-1			B=L*Y	Sb^2	tb	prob	Значимость параметров			Yf	DY^2
16	0,25	0	0		1,00484	0,001323	27,6248	3,18E-09	Параметр значим		1,00484	-2,0121	0,000257
17	0	0,25	0		-1,00255	0,001323	27,56188	3,24E-09	Параметр значим		-1,00255	-0,00699	0,000257
18	0	0	0,25		2,014386	0,001323	55,37897	1,25E-11	Параметр значим		2,014386	2,016675	0,000257
19		L=D*X^T										4,021778	0,000257
20	0,25	0,25	0,25	0,25								Дисп. адекв.	0,003088
21	0,25	-0,25	0,25	-0,25								F	0,583506
22	-0,25	-0,25	0,25	0,25								prob	0,46687
23				=ЕСЛИ(M22>$I$2;"Модель адекватно описывает систему";"Модель неадекватна эксперименту")									

Найденное значение вероятности

$$P\left(\frac{\sigma_{ad}^2}{\sigma_e^2} > \frac{s_{ad}^2}{s_e^2}\right) \approx 0{,}47$$

больше выбранного уровня значимости, поэтому гипотеза об адекватности построенной ЭС-модели экспериментальным данным не отвергается.

ПРИЛОЖЕНИЕ. Построение и анализ двухфакторной квадратичной модели с использованием программного комплекса «Градиент»

Пусть требуется построить экспериментально-статистическую модель прочности композита, получаемого совмещением матричного материала, наполнителя и модифицирующей добавки.

Модель выбрана в виде

$$R = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2,$$

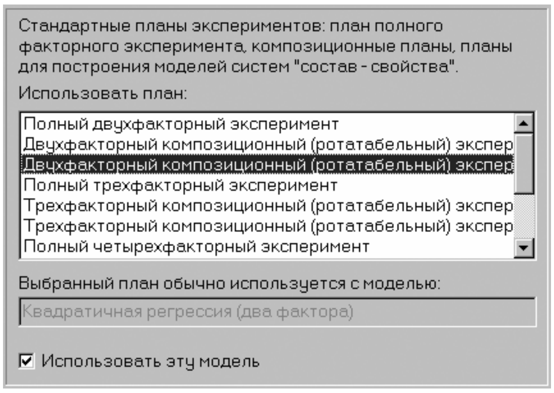
где x_1 и x_2 – нормализованные значения действующих переменных – объемной степени наполнения и концентрации добавки (в процентах от массы матричного материала).

В результате проведения натурного эксперимента по девятиточечному плану для квадратичной модели получены значения прочности, приведенные в табл. 1.

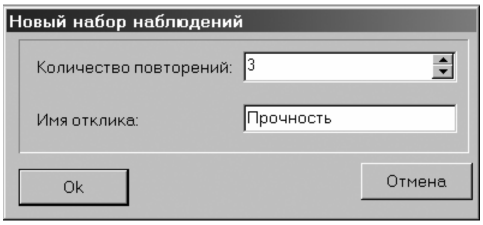
Таблица 1

Уровни входных переменных		Значения прочности в параллельных испытаниях		
X_1	X_2	R_1	R_2	R_3
0,088	0,586	71,3	69,4	68,8
0,512	0,586	108,5	110,7	112,5
0,088	3,414	87,3	85,5	83,9
0,512	3,414	124,6	125,4	123,8
0,000	2,000	58,1	60	59,6
0,600	2,000	117,6	114	114,3
0,300	0,000	96,8	100,9	98,7
0,300	4,000	120,1	115,9	117
0,300	2,000	109,8	110,5	112,8

Матрицу плана в натуральном выражении можно найти до проведения эксперимента. После запуска программы «Градиент» из меню **Файл** выбрать **Новый**. В первом из диалоговых окон создания нового плана в группе **Создать план** следует указать **Стандартный**. Во втором диалоговом окне следует выбрать девятиточечный план двухфакторного эксперимента, вместе с которым по умолчанию используется требуемая модель.



Статистический анализ предполагается выполнять по результатам параллельных опытов, число которых в каждой точке $M = 3$. Поэтому в диалоговом окне Новый набор наблюдений в поле Количество повторений следует установить значение 3. В поле Имя отклика можно ввести Прочность.



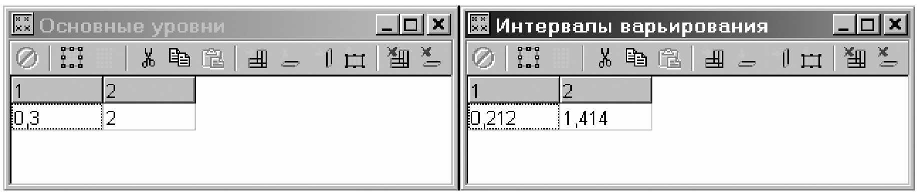
Выбранные основные уровни действующих переменных:

$$X_{1,0} = 0{,}3; \;\; X_{2,0} = 2\,\%.$$

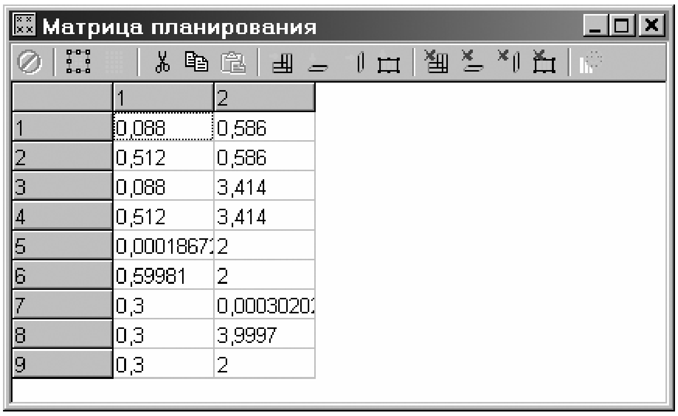
Интервалы варьирования предполагается выбрать так, чтобы минимальные значения действующих переменных были нулевыми. Так как звездное плечо выбранного плана равно $\sqrt{2}$, то интервалы варьирования:

$$\Delta X_1 = \frac{0{,}3}{\sqrt{2}} \approx 0{,}212; \;\; \Delta X_2 = \frac{2}{\sqrt{2}} \approx 1{,}414.$$

Данные значения вводятся в соответствующие поля окон Основные уровни и Интервалы варьирования.

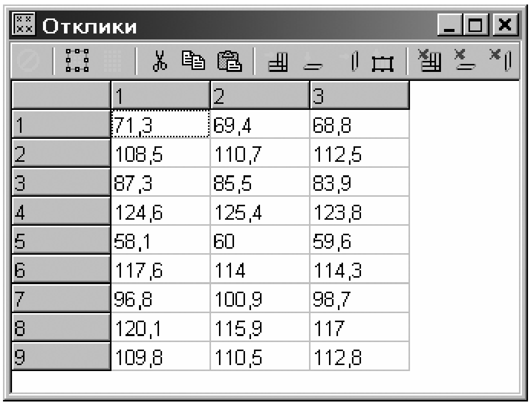


Матрицу плана в натуральном выражении Градиент создает на основе матрицы плана в кодовом выражении и указанных пользователем основных уровней и интервалов варьирования.



	1	2
1	0,088	0,586
2	0,512	0,586
3	0,088	3,414
4	0,512	3,414
5	0,0001867	2
6	0,59981	2
7	0,3	0,0003020
8	0,3	3,9997
9	0,3	2

Эмпирические значения вводятся на лист Отклики[1].

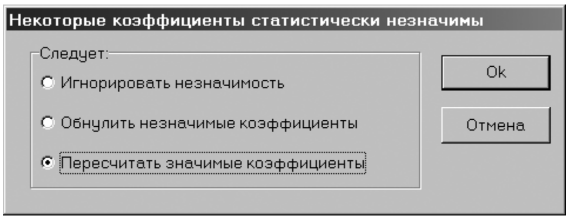


	1	2	3
1	71,3	69,4	68,8
2	108,5	110,7	112,5
3	87,3	85,5	83,9
4	124,6	125,4	123,8
5	58,1	60	59,6
6	117,6	114	114,3
7	96,8	100,9	98,7
8	120,1	115,9	117
9	109,8	110,5	112,8

Затем из меню Эксперимент выбирается Анализ. В процессе анализа имеется возможность просмотра промежуточных результатов (отображение соответствующих диалоговых окон зависит от настроек программы).

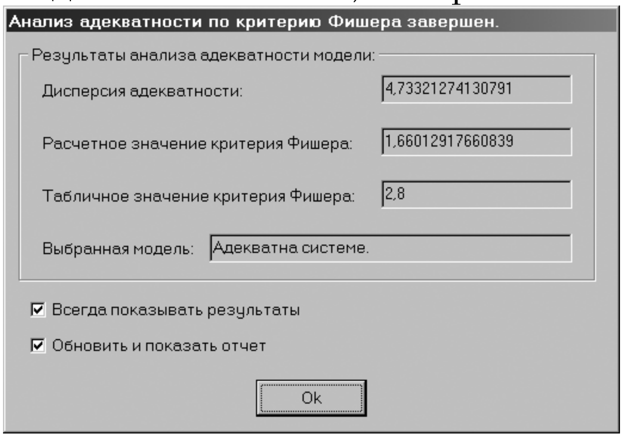
В данном примере некоторые коэффициенты ЭС-модели оказываются статистически незначимы. Поэтому на втором этапе анализа выводится диалог, позволяющий изменить вид модели.

[1] Если подлежащие анализу данные помещены на рабочий лист MS Excel, то для их копирования в таблице программы Градиент следует выделить диапазон ячеек, в который помещаются данные. Выполнить копирование значений из таблицы программы Градиент на лист MS Excel сложнее; можно предварительно *экспортировать* данные в текстовый файл.



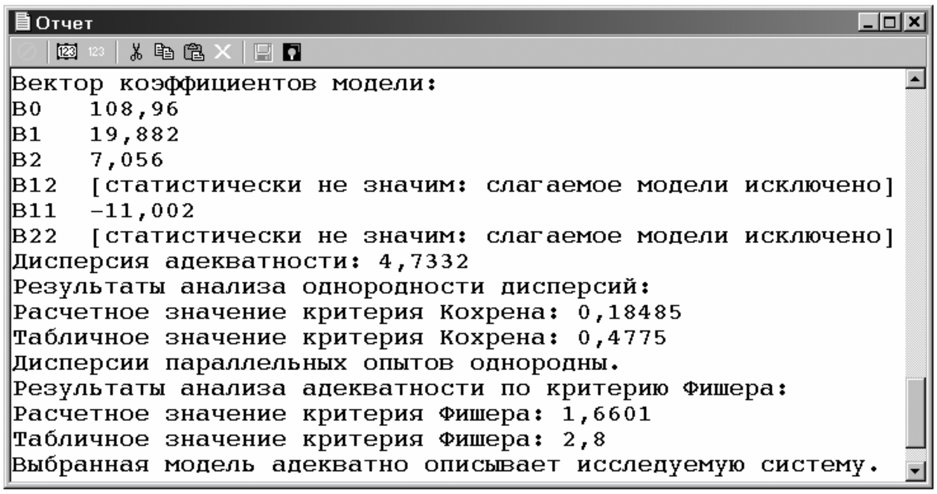
В этом диалоговом окне следует установить флажок Пересчитать значимые коэффициенты.

В последнем из диалоговых окон, отображаемых в ходе анализа



имеется флажок Обновить и показать отчет, управляющий отображением результатов. Этот флажок следует установить.

Результата анализа представляются в текстовой форме. В числе прочего среди результатов имеются значения искомых параметров модели.



Отчет

```
Вектор коэффициентов модели:
B0   108,96
B1   19,882
B2   7,056
B12  [статистически не значим: слагаемое модели исключено]
B11  -11,002
B22  [статистически не значим: слагаемое модели исключено]
Дисперсия адекватности: 4,7332
Результаты анализа однородности дисперсий:
Расчетное значение критерия Кохрена: 0,18485
Табличное значение критерия Кохрена: 0,4775
Дисперсии параллельных опытов однородны.
Результаты анализа адекватности по критерию Фишера:
Расчетное значение критерия Фишера: 1,6601
Табличное значение критерия Фишера: 2,8
Выбранная модель адекватно описывает исследуемую систему.
```

Искомая модель:

$$R = 109 + 19{,}9x_1 + 7{,}06x_2 - 11x_1^2\,.$$

В программе Градиент реализованы лишь «рудиментарные» средства визуализации данных, поэтому для построения линий равной прочности следует использовать иные средства. Поверхность отклика и изолинии прочности показаны на рис. 1.

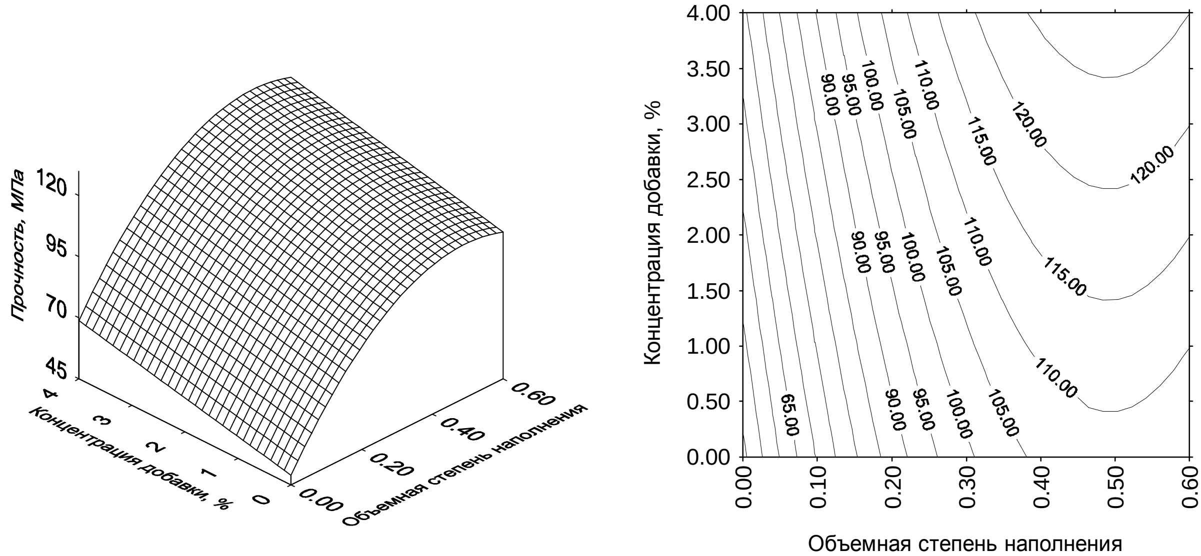


Рис. 1. Поверхность и изолинии отклика

Полученные результаты позволяют сделать ряд выводов.

Во-первых, в уравнении регрессии отсутствует слагаемое, содержащее произведение действующих переменных (коэффициент β_{12} оказался статистически незначимым). Наполнитель и модификатор действуют *независимо* друг от друга; их суммарное влияние на прочность материала оказывается равным сумме индивидуальных влияний. Значение объемной степени наполнения, соответствующее максимальной прочности, равно $X_1 = 0{,}49$ и *не зависит* от количества модификатора. Концентрация добавки рассчитывалась в процентах от массы матричного материала; поэтому можно сделать вывод о том, что модификатор изменяет свойства матрицы, не влияя на состояние межфазной границы между матрицей и наполнителем.

Во-вторых, ЭС-модель предсказывает максимальное значение прочности $R = 128$ МПа в точке $\mathbf{x} = (0{,}903; 1{,}414)$, расположенной на границе исследуемой факторной области (соответствующие значения натуральных переменных: объемная степень наполнения $X_1 = 0{,}49$, концентрация добавки $X_2 = 4\%$). Поэтому исследование нельзя считать законченным. Требуется провести дополнительный эксперимент, в ходе которого значения второго фактора будут увеличены. Далее, в ходе анализа слагаемое $\beta_{22} x_2^2$ было исключено. Линейный характер зависимости $R = R(x_2)$ свидетельствует о том, что экстремум зависимости $R = R(x_1, x_2)$ далек от точки $\mathbf{x} = (0{,}903; 1{,}414)$; по всей видимости, оптимальная концентрация модификатора существенно превышает верхний предел $X_2 = 4\%$ в поставленном эксперименте.

Библиографический список (типа)

[1]. Гмурман В.Е. Теория вероятностей и математическая статистика: учебное пособие. – М.: Высшая школа, 1977. – 575 с.
[2]. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: учебное пособие. – М.: Высшая школа, 2006 – 476 с.
[3]. Плис А.И. Mathcad: Математический практикум для инженеров и экономистов: учебное пособие. – М: Финансы и статистика, 2003 г. – 665 с.