

1. ЭКСПЕРИМЕНТ КАК ПРЕДМЕТ ИССЛЕДОВАНИЯ

1.1. Понятие эксперимента

Во многих областях научной и практической деятельности современного человека значительное место занимают теоретические методы изучения различных объектов и процессов окружающего нас мира. Так, в металлургии все более широкое применение находят результаты, полученные на основе теоретического решения задач теории пластичности, механики жидкости и газов, физики металлов, феноменологической теории разрушения материалов и других теоретических дисциплин. Однако, несмотря на высокую эффективность теоретических методов, при рассмотрении конкретных технологических проблем, особенно в условиях действующего производства, инженеру зачастую приходится сталкиваться с задачами, решение которых практически невозможно без организации и проведения того или иного экспериментального исследования.

С общепhilософской точки зрения эксперимент (от латинского *experimētum* – проба, опыт) – это чувственно-предметная деятельность в науке; в более узком смысле – опыт, воспроизведение объекта познания, проверка гипотез и т.д.

В технической литературе термину эксперимент устанавливается следующее определение – система операций, воздействий и (или) наблюдений, направленных на получение информации об объекте исследования.

Являясь источником познания и критерием истинности теорий и гипотез, эксперимент играет очень важную роль как в науке, так и в инженерной практике. Эксперименты ставятся в исследовательских лабораториях и на действующем производстве, в медицинских клиниках и на опытных сельскохозяйственных полях, в космосе и в глубинах океана.

Хотя объекты исследований очень разнообразны, методы экспериментальных исследований имеют много общего:

- каким бы простым ни был эксперимент, вначале выбирают план его проведения;
- стремятся сократить число рассматриваемых переменных, для того чтобы уменьшить объем эксперимента;
- стараются контролировать ход эксперимента;
- пытаются исключить влияние случайных внешних воздействий;
- оценивают точность измерительных приборов и точность получения данных;
- и наконец, в процессе любого эксперимента анализируют полученные результаты и стремятся дать их интерпретацию, поскольку без этого решающего этапа смысла.

К сожалению, зачастую работа экспериментатора настолько хаотична и неорганизована, а ее эффективность так мала, что полученные результаты не в состоянии оправдать даже тех средств, которые были израсходованы на проведение опытов. Поэтому вопросы организации эксперимента, снижения затрат на его проведение и обработку полученных результатов являются весьма и весьма актуальными.

Современные методы планирования эксперимента и обработки его результатов, разработанные на основе теории вероятностей и математической статистики, позволяют существенно (зачастую в несколько раз) сократить число необходимых для проведения опытов. Знание и использование этих методов делает работу экспериментатора более целенаправленной и организованной, существенно повышает как производительность его труда, так и надежность получаемых им результатов.

Как и любая другая научная дисциплина, организация и планирование эксперимента имеют свою строго определенную, во многом регламентируемую стандартами (ГОСТ 15895-77, ГОСТ 16504-81, ГОСТ 24026-80), терминологию, для первоначального знакомства с которой мы рассмотрим классификацию видов эксперимента.

1.2. Классификация видов экспериментальных исследований

Прежде всего отметим, что любой эксперимент предполагает проведение тех или иных опытов.

Опыт – воспроизведение исследуемого явления в определенных условиях проведения эксперимента при возможности регистрации его результатов.

По цели проведения и форме представления полученных результатов эксперимент делят на качественный и количественный.

Качественный эксперимент устанавливает только сам факт существования какого-либо явления, но при этом не дает никаких количественных характеристик объекта исследования. Любой эксперимент, каким бы сложным он ни был, всегда заканчивается представлением его результатов, формулировкой выводов, выдачей рекомендаций. Эта информация может быть выражена в виде графиков, чертежей, таблиц, формул, статистических данных или словесных описаний. Качественный эксперимент как раз и предусматривает именно словесное описание его результатов.

Пример 1.1. Если взять два куска стальной проволоки диаметром 5,0 мм, один из которых после пластической деформации не имел термической обработки (*проволока 5,0-II ГОСТ 3282-74*), а второй прошел операции отжига (*проволока 5,0-О-С ГОСТ 3282-74*), и подвергнуть их нескольким перегибам, то легко убедиться, что термически не обработанный металл разрушается раньше (при меньшем числе перегибов), т.е. имеет меньшую пластичность.

Однако словесное описание – не самый эффективный и информативный способ представления результатов эксперимента, поскольку он не позволяет дать количественных рекомендаций, проанализировать свойства объекта в иных условиях. Поэтому в инженерной практике основное содержание эксперимента представляется числом и количественными зависимостями.

Количественный эксперимент не только фиксирует факт существования того или иного явления, но, кроме того, позволяет установить соотношения ме-

жду количественными характеристиками явления и количественными характеристиками способов внешнего воздействия на объект исследования.

В условиях примера 1.1, для того чтобы перевести эксперимент из разряда "качественный" в "количественный", необходимо:

- определить и количественно описать те параметры процесса отжига и те свойства материала, которые по предположению могут повлиять на пластичность стали (например: температура отжига, °С; скорость охлаждения, °С/ч, фактический химический состав стали, из которой изготовлена проволока и т.д.);
- выбрать ту или иную количественную характеристику пластичности металла (например, число перегибов проволоки к моменту ее разрушения по ГОСТ 1579-93);
- в результате эксперимента необходимо установить количественную зависимость между пластичностью проволоки и параметрами процесса термообработки, с учетом тех возможных колебаний химического состава, которые допустимы для данной марки стали.

Итак, количественный эксперимент прежде всего предполагает количественное определение всех тех способов внешнего воздействия на объект исследования, от которых зависит его поведение – количественное описание всех факторов.

Фактор – переменная величина, по предположению влияющая на результаты эксперимента.

Например, в качестве факторов рассматриваемого иллюстративного эксперимента можно выбрать температуру отжига, скорость охлаждения, процентное содержание углерода или любого другого химических элемента, регламентированного для данной марки стали.

В отдельном конкретном опыте каждый фактор может принимать одно из возможных своих значений – уровень фактора.

Уровень фактора – фиксированное значение фактора относительно начала отсчета.

Например, одним уровнем такого фактора, как скорость охлаждения при отжиге, может быть 50 °С/ч, другим – 100 °С/ч и т.д.

Фиксированный набор уровней всех факторов в каждом конкретном опыте как раз и определяет одно из возможных состояний объекта исследования.

При проведении опытов очень многое зависит от того, насколько активно экспериментатор может "вмешиваться" в исследуемое явление, имеет он или нет возможность устанавливать те уровни факторов, которые представляют для него интерес.

С этой точки зрения все факторы можно разбить на три группы:

- контролируемые и управляемые – это факторы, для которых можно не только зарегистрировать их уровень, но еще и задать в каждом конкретном опыте любое его возможное значение;
- контролируемые, но неуправляемые факторы – это факторы, уровни которых можно только регистрировать, а вот задать в каждом опыте их определенное значение практически невозможно;
- неконтролируемые – это факторы, уровни которых не регистрируются экспериментатором и о существовании которых он даже может и не подозревать.

В примере 1.1 в качестве контролируемых и управляемых факторов можно очень вероятно рассматривать температуру отжига и скорость охлаждения проволоки. А вот фактическое процентное содержание различных химических элементов стали, по всей видимости, попадет в группу контролируемых, но неуправляемых факторов. Дело здесь в том, что химический состав еще может и удастся зарегистрировать (переписав его из паспорта плавки или из сопроводительных документов на данную партию проволоки), но вот задать, в условиях реального действующего сталеплавильного производства, для каждого опыта строго определенное процентное содержание, например, углерода – задача практически невыполнимая. Наконец, к группе неконтролируемых факторов в этом примере можно отнести массу причин, по которым может измениться пластичность металла (неравномерность деформации металла по длине бунта проволоки в процессе прокатки или при волочении, неблагоприятные условия

хранения металла, приводящие к его повышенной коррозии, и т.д. и т.п., на сколько в данном случае хватит фантазии исследователя).

В количественном эксперименте необходимо не только регистрировать уровни всех контролируемых факторов, но и иметь возможность устанавливать количественное описание того свойства (отклика) исследуемого явления, которое изучает (наблюдает) экспериментатор. Причем поскольку на объект исследования в процессе эксперимента всегда влияет огромное количество неконтролируемых факторов, что вносит в получаемые результаты некоторый элемент неопределенности, значение отклика, в каждом конкретном опыте, невозможно предсказать заранее. Поэтому воспроизведение исследуемого явления при одном и том же фиксированном наборе уровней всех контролируемых факторов всегда будет приводить к различным значениям отклика, т.е. отклик – это всегда случайная величина.

Отклик – наблюдаемая случайная переменная, по предположению зависящая от факторов.

Откликом в условиях примера 1.1 является пластичность стальной проволоки (количество перегибов к моменту разрушения). Причем даже если взять куски проволоки от одного и того же мотка (т.е. металл одной плавки – одинакового химического состава, имеющий один и тот же режим термообработки при одинаковой температуре отжига и скорости охлаждения), то и при этом для каждого куска проволоки мы получим разные (хотя и очень близкие друг к другу) значения пластичности металла.

И наконец, в результате количественного эксперимента необходимо найти зависимость между откликом и факторами – функцию отклика. Причем поскольку отклик – это случайная величина, то, с точки зрения теории вероятностей, его можно задать одним из параметров своего распределения, например математическим ожиданием.

Функция отклика – зависимость математического ожидания отклика от факторов.

В примере с проволокой – это зависимость математического ожидания величины пластичности стали от температуры отжига, скорости охлаждения и химического состава металла.

С учетом приведенного выше деления факторов на три группы, функцию отклика в самом общем случае можно записать в виде

$$M_y = f(x_i, h_j) + \varepsilon_\delta, \quad (1.1)$$

где M_y – математическое ожидание отклика; x_i – контролируемые и управляемые факторы; h_j – контролируемые, но неуправляемые факторы; ε_δ – ошибка эксперимента, учитывающая влияние неконтролируемых факторов.

По тому, какой группой факторов располагает исследователь, количественный эксперимент в свою очередь можно разделить еще на два вида. Если в распоряжении экспериментатора нет управляемых факторов, то такой эксперимент носит название пассивного.

Пассивный эксперимент – эксперимент, при котором уровни факторов в каждом опыте регистрируются исследователем, но не задаются.

Поскольку при пассивном эксперименте исследователь не имеет возможность задать уровень ни одного из факторов, то при проведении опытов ему остается лишь "пассивно" наблюдать за явлением и регистрировать результаты. Планирование пассивного эксперимента сводится к определению числа опытов, которые необходимо провести исследователю для решения поставленной перед ним задачи, а конечной целью пассивного эксперимента в большинстве случаев является получение функции отклика в виде

$$M_y = f(h_j) + \varepsilon_\delta. \quad (1.2)$$

Если же экспериментатор имеет возможность не только контролировать факторы, но и управлять ими, то такой эксперимент носит название активного.

Активный эксперимент – эксперимент, в котором уровни факторов в каждом опыте задаются исследователем.

Поскольку в этом случае экспериментатор имеет возможность "активно" вмешиваться в исследуемое явление, то естественно, что активный эксперимент всегда предполагает какой-либо план его проведения.

План эксперимента – совокупность данных, определяющих число, условия и порядок реализации опытов.

Поэтому активный эксперимент всегда должен начинаться с планирования.

Планирование эксперимента – выбор плана эксперимента, удовлетворяющего поставленным требованиям.

К требованиям, предъявляемым при планировании активного эксперимента, можно отнести степень точности и надежности результатов, полученных после проведения эксперимента, сроки и средства, имеющиеся в распоряжении исследователя, и т.д.

Целью активного эксперимента может быть либо определение функции отклика в виде

$$M_y = f(x_i) + \varepsilon_\delta, \quad (1.3)$$

либо поиск такого сочетания уровней управляемых факторов x_i , при котором достигается оптимальное (экстремальное – минимальное или максимальное) значение функции отклика. В этом последнем случае эксперимент носит еще название поискового (экстремального) эксперимента.

Например, если в случае с разрушением проволоки мы бы поставили перед собой целью найти такое сочетание температуры отжига и скорости охлаждения, при которых пластичность металла была бы максимальной, то наш эксперимент стал бы поисковым.

И наконец, по условиям проведения различают лабораторный и промышленный эксперименты.

Лабораторный эксперимент. В лаборатории меньше влияние случайных погрешностей, обеспечивается большая "стерильность" условий проведения опытов, в большинстве случаев осуществляется и более тщательная подготовка, одним словом, выше "культура эксперимента". Как правило, в лабораторных условиях экспериментатор может воспроизвести опыт "одинаково" значительно лучше, чем в промышленности. Это означает, что при прочих равных условиях для установления некоторого факта на заводе потребуется выполнить значительно больше опытов, чем в лаборатории. Другое важное отличие – это большая возможность варьировать (изменять) уровни факторов. Когда в лаборатории исследуется химическая реакция, температуру по желанию можно ме-

нять в широких пределах, а в металлургических печах, напротив, если ее и можно менять, то в значительно более узком диапазоне и с большей осторожностью.

Промышленный эксперимент. В промышленных условиях обеспечить условия лабораторного эксперимента значительно труднее. Усложняются измерения и сбор информации, значительно большее влияние на объект исследования и измерительные приборы оказывают различного рода помехи (резко возрастает число неконтролируемых факторов), поэтому в промышленном эксперименте особенно необходимо использовать специальные статистические методы обработки результатов. Кроме того, на реальном действующем производстве всегда желательно по возможно меньшему числу измерений получить наиболее достоверные результаты.

Контрольные вопросы

1. Что такое эксперимент? Какова его роль в инженерной практике?
2. Какие общие черты имеют научные методы исследований для изучения закономерностей различных процессов и явлений в промышленности?
3. Приведите классификации видов экспериментальных исследований, исходя из цели проведения эксперимента и формы представления результатов, а также в зависимости от условий его реализации.
4. В чем заключаются принципиальные отличия активного эксперимента от пассивного?
5. Поясните преимущества и недостатки лабораторного и промышленного эксперимента.
6. В чем отличие количественного и качественного экспериментов?
7. Дайте определения следующим терминам: опыт, фактор, уровень фактора, отклик, функция отклика, план и планирование эксперимента.

2. КРАТКИЕ СВЕДЕНИЯ ИЗ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

2.1. Случайные величины и параметры их распределений

Поскольку из-за влияния неконтролируемых факторов отклик – это всегда случайная величина, при обработке результатов эксперимента широко используется аппарат теории вероятностей и математической статистики, поэтому напомним некоторые основные понятия и определения этого раздела математики.

Случайное событие – событие, реализацию которого при определенном комплексе условий невозможно заранее предсказать.

Например, реализацию такого события, как пять остановок доменной печи в течение месяца, невозможно предсказать заранее, поскольку остановок может быть и три, и семь, и четыре, и т.д.

Случайная величина – величина, которая может принимать какое-либо значение из установленного множества и с которой связано вероятностное распределение.

Случайная величина может быть дискретной или непрерывной.

Дискретная случайная величина – случайная величина, которая может принимать значения только из конечного или счетного множества действительных чисел.

Непрерывная случайная величина – случайная величина, которая может принимать любые значения из конечного или бесконечного интервала.

Если при фиксированном наборе уровней всех контролируемых факторов провести n измерений отклика X , то в результате будет получен ряд хотя и близких, но отличающихся друг от друга значений:

$$x_i \quad (i = 1, 2, \dots, n), \quad (2.1)$$

где x_i – i -е измерение величины X ;

x_1, x_2, \dots, x_n – реализация случайной величины X .

Пример 2.1. В результате изучения работы доменной печи на протяжении полутора лет было зарегистрировано следующее количество ее остановок в течение каждого месяца (табл. 2.1).

Таблица 2.1

*Число остановок доменной печи по месяцам
(общее число наблюдений $n = 18$)*

Месяц	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Число остановок	3	4	3	5	5	5	6	4	6	5	5	2	4	6	7	5	6	7

В данном примере число остановок доменной печи в течение месяца - это дискретная случайная величина. В первом из $n = 18$ измерений этой величины было получено значение $x_1 = 3$, во втором - $x_2 = 4$ и т.д., до $x_{18} = 7$. Приведенные в табл. 2.1 значения - это реализация такой случайной величины, как число остановок доменной печи в течение месяца.

Каждому значению дискретной случайной величины X (любому из событий A , когда случайная величина X принимает какое-либо строго определенное значение x), можно поставить в соответствие следующее отношение:

$$W = \frac{m}{n}, \quad (2.2)$$

где m - число наблюдений, в которых дискретная случайная величина X оказалась равна x ; n - общее количество наблюдений. Величину W называют частотой реализации события A .

В примере 2.1, в шести наблюдениях: $i = 4, 5, 6, 10, 11$ и 16 , количество остановок доменной печи в течение месяца X оказалось равным пяти ($X = 5$), следовательно, частота реализации такого события, как пять остановок, равна $6/18 = 0,33$. Частоты реализаций для других событий (две, три, четыре и т.д. остановки) приведены в табл. 2.2.

Частота остановок доменной печи

Число остановок x	2	3	4	5	6	7
Количество наблюдений m , в которых реализовалось событие $X = x$	1	2	3	6	4	2
Частота реализации, $W = m / n$	0,06	0,11	0,17	0,33	0,22	0,11

Если продолжить наблюдения за работой доменной печи в течение еще полутора лет, то, конечно же, совершенно не обязательно, что на протяжении следующих восемнадцати месяцев пять остановок будет снова зарегистрировано ровно в 6 случаях из 18 наблюдений, а частота реализации этого события опять окажется равной 0,33. Однако при возрастании числа повторений одного и того же комплекса условий частота реализации такого события, как, например, пять остановок печи в течение месяца, будет принимать все более и более устойчивое значение. Так, если подсчитать частоту реализации данного события за 36 месяцев, то она уже практически не будет отличаться от того значения, которое затем можно будет получить за четыре с половиной года (при условии, что за все это время наблюдений в работе доменной печи не произойдет никаких существенных изменений).

Предел, к которому стремится отношение m/n при неограниченном возрастании числа опытов n , называется вероятностью случайного события.

Вероятность $P(A)$ события A – число от нуля до единицы, которое представляет собой предел частоты реализации события A при неограниченном числе повторений одного и того же комплекса условий.

Для дискретной случайной величины можно указать вероятность, с которой она принимает каждое из своих возможных значений конечного или счетного множества действительных чисел. Для непрерывной случайной величины задают вероятность ее попадания в один из заданных интервалов области ее определения (поскольку вероятность того, что она примет какое-либо конкретное свое значение, стремится к нулю).

Полностью свойства случайной величины описываются законом ее распределения, под которым понимают связь между возможными значениями случайной величины и соответствующими им вероятностями.

Распределение случайной величины – функция, которая однозначно определяет вероятность того, что случайная величина принимает заданное значение или принадлежит к некоторому заданному интервалу.

В математике используют два способа описания распределений случайных величин: интегральный (функция распределения) и дифференциальный (плотность распределения).

Функция распределения $F(x)$ – функция, определяющая для всех действительных x вероятность того, что случайная величина X принимает значение не больше, чем x .

$$F(x) = P(X \leq x). \quad (2.3)$$

Функция распределения $F(x)$ имеет следующие свойства (рис.2.1, а):

1. Ее ордината, соответствующая произвольной точке x_1 , представляет собой вероятность того, что случайная величина X будет меньше, чем x_1 , т.е. $F(x_1) = P(X \leq x_1)$.
2. Функция распределения принимает значение, заключенное между нулем и единицей:

$$0 \leq F(x) \leq 1. \quad (2.4)$$

3. Функция распределения стремится к нулю при неограниченном уменьшении x и стремится к единице при неограниченном возрастании x , то есть

$$\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow +\infty} F(x) = 1. \quad (2.5)$$

4. Функция распределения представляет собой монотонно возрастающую кривую, то есть

$$F(x_2) > F(x_1), \text{ если } x_2 > x_1. \quad (2.5a)$$

5. Ее приращение на произвольном отрезке $(x_1; x_2)$ равно вероятности того, что случайная величина X попадет в данный интервал:

$$F(x_2) - F(x_1) = P(X \leq x_2) - P(X \leq x_1) = P(x_1 < X \leq x_2). \quad (2.6)$$

Рассмотрим, какие особенности имеют функции распределения дискретных случайных величин. Пусть X – дискретная случайная величина, принимающая возможные значения x_1, x_2, \dots, x_n с вероятностями p_1, p_2, \dots, p_n . Функция распределения вероятностей этой случайной величины X равна

$$F(x) = P(X \leq x) = \sum_{x_k} p_k,$$

где производится суммирование вероятностей всех возможных значений случайной величины X , меньших чем x . Такая функция всегда разрывная, ступенчатая (рис.2.1, б): от $-\infty$ до x_1 включительно функция равна нулю, в точке x_1 происходит скачок на величину p_1 , и функция остается постоянной до x_2 включительно и т.д., то есть возможным значениям случайной величины соответствуют скачки функции, равные вероятностям этих значений. Последний скачок на p_n происходит в точке x_n , и функция равна единице от x_n до $+\infty$. Таким образом, сумма всех скачков равна единице.

Плотность распределения $f(x)$ – первая производная (если она существует) функции распределения.

$$f(x) = \frac{dF(x)}{dx}. \tag{2.7}$$

Плотность функции распределения $f(x)$ имеет следующие свойства (рис.2.2):

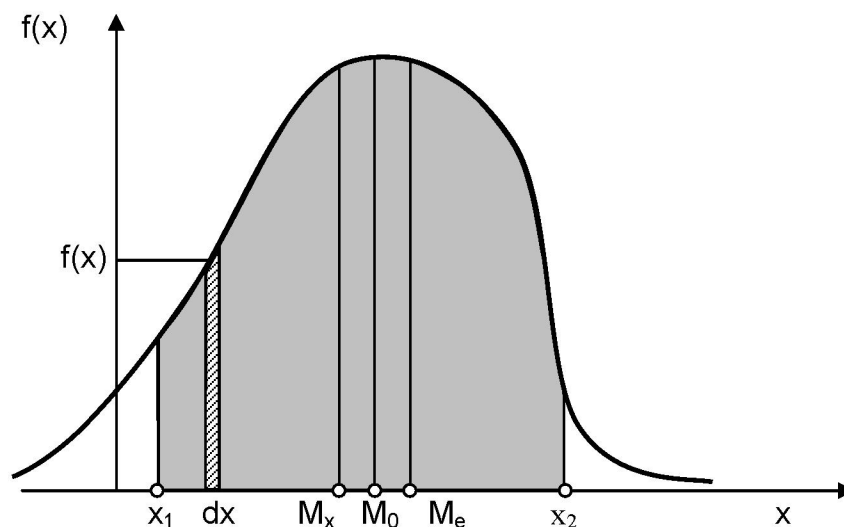


Рис.2.2. Дифференциальный закон распределения – плотность распределения $f(x)$

1. Плотность распределения вероятностей является неотрицательной функцией, т.е.

$$f(x) \geq 0. \quad (2.8)$$

Это свойство справедливо, так как $F(x)$ есть неубывающая функция.

2. Функция распределения случайной величины X равна определенному интегралу от плотности распределения вероятностей в пределах $(-\infty, x)$:

$$F(x) = \int_{-\infty}^x f(x) dx. \quad (2.9)$$

3. Вероятность события, состоящая в том, что случайная величина X примет значение, заключенное в полуинтервале $[x_1, x_2]$, равна определенному интегралу от плотности распределения вероятностей на этом полуинтервале:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x) dx. \quad (2.10)$$

4. Интеграл плотности распределения в бесконечно большом интервале $(-\infty, +\infty)$ равен единице:

$$\int_{-\infty}^{+\infty} f(x) dx = P(-\infty < X \leq +\infty) = 1, \quad (2.11)$$

так как попадание случайной величины в интервал $-\infty < X < +\infty$ есть достоверное событие.

В большинстве случаев при обработке экспериментальных данных, основываясь на тех или иных предположениях (гипотезах) относительно свойств исследуемой случайной величины, удается записать функцию ее распределения (а следовательно, и плотность распределения как первую производную от функции распределения) с точностью до некоторых неизвестных параметров.

Например, для случайной величины, которая удовлетворяет так называемому нормальному закону распределения (закону распределения Гаусса), функцию распределения можно записать в виде

$$F(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^x e^{-\frac{(x-M_x)^2}{2\sigma_x^2}} dx, \quad (2.12)$$

а для случайной величины, имеющей, например, распределение Вейбулла-Гнеденко (используемое для описания результатов экспериментов в случае хрупкого разрушения металла, а также в испытаниях на многоцикловую усталость), функция распределения определяется следующим выражением:

$$F(x) = 1 - e^{-\left(\frac{x-x_H}{c}\right)^b}, \quad \text{при } X > x_H,$$

$$F(x) = 0, \quad \text{при } X \leq x_H. \quad (2.13)$$

В функциях (2.12) и (2.13) константы M_x , σ_x^2 и c , b , x_H являются параметрами распределений, причем первое из этих двух выражений относится к двухпараметрическому виду закона распределения, а второе, соответственно, – к трехпараметрическому.

Параметр распределения – постоянная, от которой зависит функция распределения.

Следовательно, если известен вид функции распределения (каким-либо образом установлено, что случайная величина не противоречит тому или иному закону распределения), то для того, чтобы однозначно охарактеризовать случайную величину, достаточно задать только лишь параметры ее распределения.

Важнейшими параметрами распределения, задающими случайную величину X , являются ее математическое ожидание M_x (характеризует центр рассеивания) и дисперсия σ_x^2 (характеризует степень рассеивания).

Математическое ожидание M_x – среднее взвешенное по вероятностям значение случайной величины.

Для дискретной случайной величины математическое ожидание определяется выражением

$$M_x = \sum_i x_i p_i, \quad (2.14)$$

где x_i – значения дискретной случайной величины, а $p_i = P(X = x_i)$.

Если в условиях примера 2.1 предположить, что $p_i \approx W_i$ (см. табл. 2.2), то для математического ожидания такой дискретной случайной величины, как число остановок доменной печи в течение месяца, можно получить следующее значение:

$$M_x = 2 \cdot 0,06 + 3 \cdot 0,11 + 4 \cdot 0,17 + 5 \cdot 0,33 + 6 \cdot 0,22 + 7 \cdot 0,11 = 4,87.$$

Для непрерывной случайной величины математическое ожидание определяется интегралом

$$M_x = \int_{-\infty}^{+\infty} x f(x) dx, \quad (2.15)$$

где $f(x)$ – плотность распределения непрерывной случайной величины.

Можно отметить, что геометрический смысл математического ожидания непрерывной случайной величины – это абсцисса центра тяжести фигуры под кривой плотности распределения $f(x)$. Сказанное проиллюстрируем на рис. 2.2, где видно, что произведение $f(x)dx$ есть площадь элементарного участка под кривой $f(x)$, а x – абсцисса этого участка, т.е. расстояние от начала координат.

Следовательно, интеграл (2.15) дает абсциссу центра тяжести всей площади фигуры под кривой $f(x)$.

Кроме математического ожидания центр рассеивания случайной величины можно еще охарактеризовать такими параметрами ее распределения, как мода и медиана.

Мода M_o – значение случайной величины, соответствующее локальному максимуму плотности вероятностей для непрерывной случайной величины или локальному максимуму вероятности для дискретной случайной величины.

Для примера 2.1 (см. табл. 2.2), при условии, что $p_i \approx W_i$, мода M_o числа остановок доменной печи равна 5, поскольку именно этому значению данной дискретной случайной величины соответствует локальный максимум вероятности, равный 0,33.

Медиана M_e – значение случайной величины, для которого функция распределения принимает значение $\frac{1}{2}$, или имеет место «скачок» со значения, меньшего чем $\frac{1}{2}$, до значения, большего чем $\frac{1}{2}$.

Таким образом, для дифференциального закона распределения медиана есть такое значение непрерывной случайной величины X , которое делит пополам площадь под кривой плотности распределения $f(x)$.

В примере 2.1, если предположить, что функция распределения от четырех остановок $F(4)$ (вероятность того, что число остановок доменной печи в течение месяца будет не более четырех) равна $0,06 + 0,11 + 0,17 = 0,34$, а функция распределения $F(5) = 0,34 + 0,33 = 0,67$, то медианой M_e такой дискретной случайной величины, как число остановок доменной печи в течение месяца, будет значение $M_e = 5$.

Дисперсия случайной величины σ_x^2 – математическое ожидание случайной величины $(X - M_x)^2$.

Для дискретной случайной величины дисперсия определяется следующим математическим выражением:

$$\sigma_x^2 = \sum_{i=1}^n (x_i - M_x)^2 \cdot p(x_i). \quad (2.16)$$

В примере 2.1 (опять же, если предположить, что $p_i \approx W_i$) значение дисперсии числа остановок доменной печи равно:

$$\sigma_x^2 = (2 - 4,87)^2 \cdot 0,06 + (3 - 4,87)^2 \cdot 0,11 + (4 - 4,87)^2 \cdot 0,17 +$$

$$+ (5 - 4,87)^2 \cdot 0,33 + (6 - 4,87)^2 \cdot 0,22 + (7 - 4,87)^2 \cdot 0,11 = 1,7931.$$

Для непрерывной случайной величины дисперсия определяется выражением

$$\sigma_x^2 = \int_{-\infty}^{+\infty} (x - M_x)^2 \cdot f(x) dx, \quad (2.17)$$

где x – значения непрерывной случайной величины X ; $f(x)$ – плотность распределения; M_x – математическое ожидание.

Дисперсия имеет размерность квадрата единицы измерения случайной величины, а положительное значение квадратного корня из дисперсии называется средним квадратичным отклонением.

Среднее квадратичное отклонение σ_x – неотрицательный квадратный корень из дисперсии.

$$\sigma_x = +\sqrt{\sigma_x^2}. \quad (2.18)$$

Для примера 2.1 среднее квадратичное отклонение числа остановок доменной печи в течение месяца равно $\sigma_x = +\sqrt{1,7931} = 1,34$.

В заключение этого раздела дадим определение еще одного параметра распределения случайной величины, который носит название квантиль.

Квантиль порядка P , x_p – значение случайной величины, для которого функция распределения принимает значение P или имеет место «скачок» со значения, меньшего чем P , до значения, большего чем P :

$$F(x_p) = P. \quad (2.19)$$

Из этого определения квантиля следует, что медиана Me – это квантиль порядка $1/2$, т.е. $Me = x_{0,5}$.

Вероятность попадания случайной величины X в интервал $[x_{P1}, x_{P2}]$ равна

$$P(x_{P1} < X \leq x_{P2}) = P(X \leq x_{P2}) - P(X \leq x_{P1}) = F(x_{P2}) - F(x_{P1}) = P_2 - P_1. \quad (2.20)$$

В примере 2.1 квантиль порядка 0,95 числа остановок доменной печи скорее всего равен семи $x_{0,95} = 7$, поскольку $F(6) \approx 0,06 + 0,11 + 0,17 + 0,33 + 0,22 = 0,89$, а $F(7) \approx 0,89 + 0,11 = 1,00$.

2.2. Нормальный закон распределения

Функция распределения $F(x)$ и соответствующая ей плотность распределения $f(x)$ представляют собой некоторую математическую модель свойств исследуемой случайной величины (отклика), значения которой регистрируются в ходе эксперимента. Поэтому одной из основных задач статистической обработки опытных данных является нахождение таких функций распределения, которые, с одной стороны, достаточно хорошо описывали бы наблюдаемые значения случайной величины, а с другой – были бы удобны для дальнейшего статистического анализа. При этом вид функции распределения предпочтительно выбирать на основе представлений о физической природе рассматриваемого явления, т.к. в этом случае исключаются возможные погрешности при распространении найденных закономерностей за пределы изучаемого в эксперименте интервала варьирования (изменения) случайной величины (отклика).

Из всех изученных к настоящему времени случайных величин при обработке экспериментальных данных исследователи чаще всего оперируют со случайными величинами, которые имеют так называемое нормальное (Гауссово) распределение (рис.2.3). Не вдаваясь в подробные математические выкладки, отметим, что, согласно *центральной предельной теореме* математической статистики, «при определенных условиях распределение нормированной суммы n независимых случайных величин, распределенных по произвольному закону, стремится к нормальному, когда n стремится к бесконечности». Необходимые условия, при которых эта теорема оказывается справедливой, состоят в том, что различные случайные величины должны иметь конечные дисперсии и дисперсия любой случайной величины не должна быть слишком большой по сравнению с дисперсиями других.

При обработке экспериментальных данных эта теорема имеет очень большое значение, поскольку отклик становится случайной величиной в результате влияния неконтролируемых факторов, число которых скорее всего стремится к бесконечности. Кроме того, если при проведении опытов все наиболее существенные факторы контролируются, то воздействие на отклик каждого из неконтролируемых факторов не должно быть слишком большим по сравнению с остальными неконтролируемыми факторами. Другими словами, та дисперсия (рассеивание) отклика, которую вызывает какой-либо из неконтро-

лируемых факторов, не должна сильно отличаться от дисперсий, связанных с влиянием остальных неконтролируемых факторов. В противном случае фактор, дисперсия от которого существенно отличается от других, обязательно должен быть переведен в разряд контролируемых.

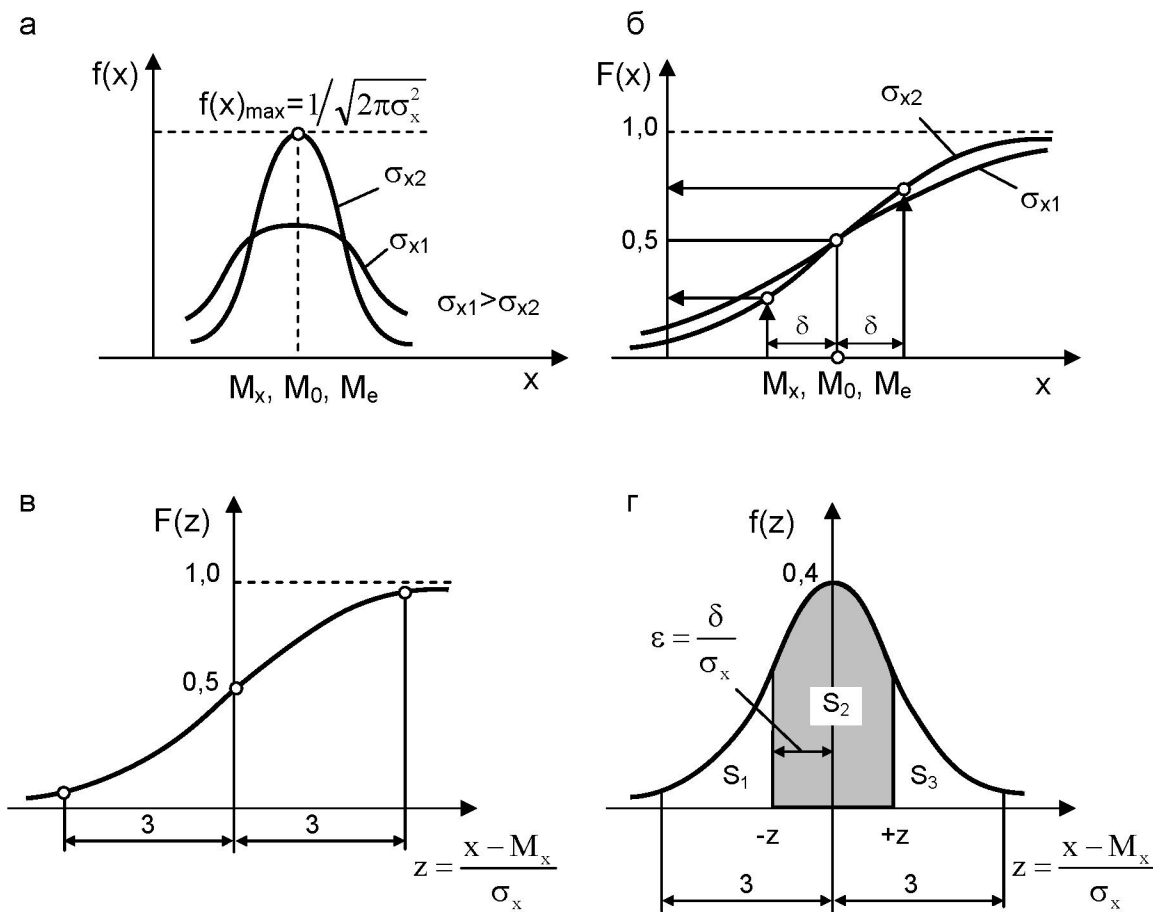


Рис.2.3. Плотность распределения (а,г) и функция распределения (б,в) при нормальном законе распределения случайных величин

Следовательно, если при планировании эксперимента учтены все наиболее существенные факторы и затем, при проведении опытов, они контролируются, то при обработке экспериментальных данных можно предполагать, что отклик не должен противоречить нормальному распределению.

Как правило, нормальному закону подчиняются результаты испытаний стали на прочность, производительность многих металлургических агрегатов, составы сырья, топлива, сплавов, массы слитков, отлитых в однотипные изложницы, случайные ошибки измерений и т.п., поэтому при обработке результатов наблюдений исследователи прежде всего предполагают именно нормальное распределение отклика.

Большинство других распределений, которые используются в математической статистике (Стьюдента, Фишера, Пирсона, Кохрена, а также распределения, по которым составлены различные критериальные таблицы), получены на основе нормального распределения.

Нельзя, однако, абсолютизировать значение нормального распределения. Не все случайные величины распределены по нормальному закону. Тем не менее на практике, если явление подвержено действию многих случайных факторов, их суммарное воздействие вполне оправданно можно описать с помощью нормального закона.

Как уже было отмечено, для случайной величины, которая не противоречит нормальному закону, функция распределения (2.12) и соответствующая ей плотность распределения

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \cdot e^{-\frac{[x-M_x]^2}{2\sigma_x^2}} \quad (2.21)$$

определяются двумя параметрами: M_x – математическим ожиданием и σ_x^2 – дисперсией.

Отметим некоторые свойства нормального закона распределения.

1. Кривая плотности распределения симметрична относительно значения M_x , называемого иногда центром распределения.

2. При бóльших значениях σ_x^2 кривая $f(x)$ более пологая, т.е. σ_x^2 является мерой величины рассеивания значения случайной величины около значений M_x . При уменьшении параметра σ_x^2 кривая нормального распределения сжимается вдоль оси ОХ и вытягивается вдоль $f(x)$.

3. Максимум ординаты кривой плотности распределения определяется выражением

$$f_{\max} = \frac{1}{\sqrt{2\pi\sigma_x^2}}, \quad (2.22)$$

что при $\sigma_x^2=1$ соответствует значению примерно 0,4.

4. Для нормального распределения математическое ожидание, мода и медиана совпадают:

$$M_x = M_o = M_e. \quad (2.23)$$

В ряде случаев рассматривается не сама случайная величина X , а ее отклонение от математического ожидания:

$$Y = X - M. \quad (2.24)$$

Такая случайная величина Y называется центрированной.

Отношение случайной величины X к ее среднему квадратичному отклонению

$$V = \frac{X}{\sigma_x} \quad (2.25)$$

называется нормированной случайной величиной.

Таким образом, центрированная случайная величина – разность между данной случайной величиной и ее математическим ожиданием, а нормированная случайная величина – отношение данной случайной величины к ее среднему квадратичному отклонению.

Очевидно, что математическое ожидание центрированной случайной величины равно нулю, $M_y = 0$, а дисперсия нормированной случайной величины равна единице, $\sigma_v^2 = 1$.

Приведенная случайная величина – центрированная и нормированная случайная величина

$$Z = \frac{X - M_x}{\sigma_x}. \quad (2.26)$$

Математическое ожидание и дисперсия приведенной случайной величины Z равны соответственно нулю, $M_z = 0$, и единице, $\sigma_z^2 = 1$.

Нормальное распределение с параметрами $M_z = 0$ и $\sigma_z^2 = 1$ называется стандартным (нормированным).

Для приведенной случайной величины нормальное стандартное распределение принимает вид

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{z^2}{2}} dz = \Phi(z), \quad (2.27)$$

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}} = \phi(z). \quad (2.28)$$

Графики этих функций показаны на рис. 2.3 в, г, причем

$$\Phi(-z) = 1 - \Phi(z), \quad (2.29)$$

$$\phi(-z) = \phi(z). \quad (2.30)$$

Покажем справедливость соотношения (2.29). Рассмотрим график плотности стандартного нормального распределения (см. рис. 2.3,г). Обозначим площадь под ним левее точки $-z$ через S_1 ; площадь между $-z$ и z – через S_2 , оставшуюся площадь (правее z) – через S_3 . Тогда, во-первых, из симметричности графика плотности следует, что $S_1 = S_3$. Во-вторых, $S_1 + S_2 + S_3 = 1$ или $S_1 + (S_1 + S_2) = 1$ (вся площадь под графиком плотности равна единице). По смыслу функции распределения $S_1 = \Phi(-z)$, $S_1 + S_2 = \Phi(z)$. Следовательно, $\Phi(-z) + \Phi(z) = 1$, откуда и следует равенство (2.29).

Значения нормированной функции (2.27) нормального распределения (функции Лапласа) и значения плотности нормированного нормального распределения (2.28) табулированы и приведены в различных учебниках и справочниках по математической статистике (наиболее подробные таблицы см. [11]). В списке статистических функций электронных таблиц Microsoft Excel им соответствуют **НОРМРАСП**(x ; 0; 1; ИСТИНА) или **НОРМСТРАСП**(z) – для (2.27) и **НОРМРАСП**(x ; 0; 1; ЛОЖЬ) – для (2.28).

Геометрически функция Лапласа представляет площадь под кривой $f(z)$ в интервале от $-\infty$ до некоторой конкретной величины z .

Заметим, что иногда вместо функции $\Phi(z)$ табулируется функция $\Phi_0(z)$:

$$\Phi_0(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{z^2}{2}} dz,$$

равная площади под графиком стандартного нормального распределения от 0 до z (см. рис.2.3, г).

В силу симметрии

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{z^2}{2}} dz = 1/2.$$

Поэтому между функциями и существует простая зависимость

$$\Phi(z) = 1/2 + \Phi_0(z).$$

Функция $\Phi_0(z)$ нечетна:

$$\Phi_0(-z) = -\Phi_0(z).$$

В самом деле,

$$\Phi_0(-z) = \Phi(-z) - 1/2 = 1 - \Phi(z) - 1/2 = 1/2 - (1/2 + \Phi_0(z)) = -\Phi_0(z).$$

В соответствии с (2.19) квантиль z_p порядка p , нормированного нормального закона распределения - это такое значение приведенной случайной величины Z , для которого функция распределения (2.27) принимает значение P :

$$\Phi(z_p) = P. \tag{2.31}$$

При определении квантили z_p необходимо решать задачу, обратную задаче определения значений функции Лапласа, т.е. по известному значению P этой функции (2.27) находить соответствующее ему значение аргумента z_p . Для этого можно либо воспользоваться табулированными значениями функции Лапласа (например, поскольку $\Phi(1,64) = 0,94950$, а $\Phi(1,65) = 0,9505$, то $z_{0,95} \approx 1,645$), либо воспользоваться таблицами для функции, обратной функции Лапласа, т.е. табулированными значениями квантилей нормированного нормального закона распределения (см. [11] или приложение). Определение квантили z_p в электронных таблицах Microsoft Excel сводится к вычислению статистической функции **НОРМОБР**(P ; 0; 1) или **НОРМСТОБР**(P) (например, **НОРМОБР**(0,95; 0; 1) = **НОРМСТОБР**(0,95) = 1,644853).

Для квантили стандартного нормального распределения справедливо следующее равенство:

$$z_{1-p} = -z_p. \tag{2.32}$$

Рассмотрим график плотности стандартного нормального распределения (рис.2.4). Площадь под графиком левее квантили z_p по определению равна p . Значит, площадь правее этой точки равна $1 - p$. Такая же площадь расположена левее точки z_{1-p} . Итак, площади левее z_{1-p} и правее z_p равны. Поскольку

график симметричен относительно оси ординат, из этого следует, что эти точки расположены на одинаковом расстоянии от нуля.

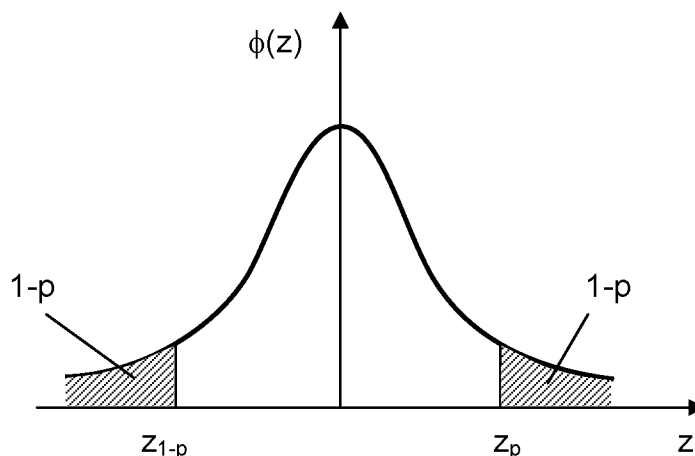


Рис. 2.4. Квантиль стандартного нормального распределения

Зная квантиль z_p порядка p нормированного нормального закона распределения ($M_z = 0$ и $\sigma_z^2 = 1$), всегда можно найти квантиль x_p соответствующего порядка p для нормального распределения с произвольными параметрами M_x и σ_x^2 .

Поскольку

$$F(x_p) = P(X \leq x_p) = P\left(\frac{X - M_x}{\sigma_x} \leq \frac{x_p - M_x}{\sigma_x}\right) = P\left(Z \leq \frac{x_p - M_x}{\sigma_x}\right) = \Phi\left(\frac{x_p - M_x}{\sigma_x}\right) = P = \Phi(z_p),$$

то $\frac{x_p - M_x}{\sigma_x} = z_p$ и, следовательно,

$$x_p = M_x + z_p \sigma_x. \quad (2.32a)$$

В ряде случаев важно знать вероятность того, что случайная величина X , подчиняющаяся нормальному закону распределения, не будет отличаться от своего математического ожидания M_x больше чем на величину $\pm \delta = \varepsilon \cdot \sigma_x$ (см. рис. 2.3, г).

$$\begin{aligned} P(M_x - \delta < X \leq M_x + \delta) &= P\left(\frac{M_x - \varepsilon \sigma_x - M_x}{\sigma_x} < \frac{X - M_x}{\sigma_x} \leq \frac{M_x + \varepsilon \sigma_x - M_x}{\sigma_x}\right) = \\ &= P(-\varepsilon < Z \leq +\varepsilon) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\varepsilon} e^{-\frac{z^2}{2}} dz - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\varepsilon} e^{-\frac{z^2}{2}} dz = \Phi(\varepsilon) - \Phi(-\varepsilon) = \end{aligned}$$

$$= \Phi(\varepsilon) - (1 - \Phi(\varepsilon)) = 2\Phi(\varepsilon) - 1. \quad (2.33)$$

Так, при $\delta = \sigma_x$ ($\varepsilon = 1$) получаем, что $P(M_x + \sigma_x < X \leq M_x - \sigma_x) = 2\Phi(1) - 1$, а поскольку по таблицам (например, см. табл. [11]) $\Phi(1) = 0,84135$ (или в Microsoft Excel **НОРМРАСП(1;0;1;ИСТИНА) = НОРМСТРАСП(1) = 0,84135**), то для случайной величины с нормальным законом распределения вероятность того, что она примет такое значение, которое не будет отличаться от ее математического ожидания более чем на одно среднее квадратическое отклонение, равна $2 \cdot 0,84135 - 1 = 0,68$. Иными словами, при нормальном распределении примерно 2/3 всех значений случайной величины (отклика) лежит в интервале $M_x \pm \sigma_x$.

Аналогично можно подсчитать, что интервалу $M_x \pm 1,96\sigma_x \approx M_x \pm 2\sigma_x$ соответствует вероятность 0,95 ($\Phi(1,96) = 0,975002$), а интервалу $M_x \pm 3\sigma_x$ - 0,997 ($\Phi(3) = 0,99865$). Отметим дополнительно, что 90% значений случайной величины лежат в диапазоне $M_x \pm 1,64\sigma_x$ ($\Phi(1,64) = 0,949497$).

Следовательно, отличие какого-либо из значений случайной величины с нормальным законом распределения от ее математического ожидания не превосходит утроенного среднего квадратического отклонения с вероятностью 0,997. Это свойство в математической статистике носит название «правило трех сигм».

Чем больше величина интервала $M_x \pm \delta$, тем с большей вероятностью случайная величина X попадает в этот интервал.

Рассмотрим небольшой пример.

Пример 2.2. Предположим, что математическое ожидание содержания кремния в чугуне равно $M_{Si} = 0,6\%$, а среднеквадратичное отклонение $\sigma_{Si} = 0,15\%$. В этом случае мы можем быть уверены в том, что величина фактически измеренного значения процентного содержания кремния в чугуне будет находиться в интервалах:

$$0,6 \pm 1,00 \cdot 0,15 = 0,6 \pm 0,15 \text{ с вероятностью } 0,68;$$

$$0,6 \pm 1,64 \cdot 0,15 = 0,6 \pm 0,25 \text{ с вероятностью } 0,90;$$

$$0,6 \pm 1,96 \cdot 0,15 = 0,6 \pm 0,29 \text{ с вероятностью } 0,95;$$

$$0,6 \pm 3,00 \cdot 0,15 = 0,6 \pm 0,45 \text{ с вероятностью } 0,997,$$

т.е. из 1000 проб только 3 пробы по содержанию кремния в чугуне будут выходить из диапазона от 0,15 до 1,05%.

Заметим, однако: при рассмотрении примера 2.2 мы предполагали, что процентное содержание кремния в чугуне не противоречит нормальному закону распределения, а также то, что нам изначально были известны математическое ожидание M_x и среднеквадратичное отклонение σ_x этой случайной величины, т.е. было выполнено большое (в пределе бесконечное) число измерений. Как же работать со случайными величинами в реальных условиях проведения эксперимента, когда число измерений весьма ограничено? К рассмотрению методологии решения подобных задач мы и перейдем в следующем разделе.

3. ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Предварительная обработка результатов измерений и наблюдений необходима для того, чтобы в дальнейшем, при построении эмпирических зависимостей (функций отклика), с наибольшей эффективностью использовать статистические методы и корректно анализировать полученные результаты.

Содержание предварительной обработки состоит в отсеивании грубых погрешностей и оценке достоверности результатов измерений. Другими важными моментами предварительной обработки данных являются проверка соответствия результатов измерения нормальному закону и определение параметров этого распределения. Если гипотеза о том, что отклик не противоречит нормальному распределению, окажется неприемлемой, то следует определить, какому закону распределения подчиняются опытные данные или, если это возможно, преобразовать опытное распределение к нормальному виду.

3.1. Вычисление параметров эмпирических распределений. Точечное оценивание

Рассмотрение вопросов обработки экспериментальных данных начнем с простейшей ситуации, когда отклик регистрируется при фиксированных уровнях всех контролируемых факторов и при проведении опытов (в результате влияния неконтролируемых факторов) исследователь получает хотя и близкие, но отличные друг от друга результаты.

Пример 3.1. При производстве железнодорожных рельсов широкой колеи типа Р65 (по ГОСТ 18267-82) были получены следующие три значения твердости *HV* (по ГОСТ 9012-59) на поверхности катания головки одного и того же рельса (на обоих концах на расстоянии не более 1 м от торцов и в средней части рельса): 351, 370 и 365.

Попытаемся найти ответ на вопрос – чему равна твердость на поверхности катания данного рельса?

На первый взгляд решение поставленной задачи не вызывает никаких особых проблем, и начинающие исследователи, не особенно искушенные в области теории вероятностей и математической статистики, скорее всего ответят, что твердость на поверхности катания рельса равна (HB_1 – первый вариант ответа)

$$HB_1 = (351 + 370 + 365)/3 = 362,00 ,$$

т.е. будет найдено среднее арифметическое (выборочное среднее арифметическое) из трех полученных значений отклика.

Однако опытные данные можно усреднять и другими способами. Например, можно подсчитать среднее геометрическое (HB_2 – второй вариант ответа):

$$HB_2 = \sqrt[3]{351 \cdot 370 \cdot 365} = 361,91,$$

или найти среднее, только между минимальным (351) и максимальным (370) значениями – так называемую середину размаха (HB_3 – третий вариант ответа):

$$HB_3 = (351 + 370)/2 = 360,50,$$

или, расположив все значения в возрастающей последовательности 351, 365, 370, взять средний член полученного ряда – средний член вариационного ряда (HB_4 – четвертый вариант ответа):

$$HB_4 = 365,00.$$

Можно придумать и какие-либо другие способы (например, очень «оригинальной» может быть идея еще раз усреднить все четыре полученных значения), однако остановимся пока только на этих четырех вариантах ответа на поставленный перед нами вопрос.

Мы видим, что, не привлекая никаких дополнительных соображений, нам пока достаточно трудно обосновать тот или иной вариант, на котором было бы предпочтительно остановиться.

Так, если выбирать тот ответ, который потребует от нас меньшего количество вычислений, то тогда лучше всего отдать предпочтение значению $HB_4=365,00$ (вообще не требует никаких расчетов). Однако подобное обоснование вряд ли можно считать достаточно надежным и убедительным.

Поэтому давайте остановимся и задумаемся о том, почему вообще мы столкнулись с подобной ситуацией.

Ведь если бы, например, нам нужно было найти ответ на вопрос, какое количество проходов при прокатке данного профиля осуществляется в двух-валковых рельсовых калибрах, и мы по ходу технологического процесса проследили за тремя различными раскатами, то в результате было бы получено три абсолютно одинаковых значения (допустим, пять). В подобной ситуации нет необходимости считать ни выборочное среднее, ни среднее геометрическое, ни середину размаха, ни находить средний член вариационного ряда и т.д., поскольку можно сразу указать то количество рельсовых калибров, которые проходит раскат в процессе прокатки.

Следовательно, между такими величинами, как число рельсовых калибров и твердость на поверхности катания головки, есть принципиальная разница, которая заключается в том, что первая из двух названных величин является детерминированной, а вторая – случайной. И если для того, чтобы описать детерминированную величину, достаточно указать одно ее значение (например, число рельсовых калибров равно пяти), то для описания случайной величины нужно знать ее распределение. Другими словами, для случайной величины недостаточно указать только лишь какое-либо ее значение (или комбинацию ее значений, как, например, выборочное среднее арифметическое), а нужно записать функцию, которая однозначно определяет вероятность того, что случайная величина принимает заданное значение или принадлежит к некоторому заданному интервалу.

Поэтому ответ на вопрос примера 3.1 надо начинать не с поиска каких-либо вариантов усреднения опытных данных, а прежде всего с констатации того факта, что твердость на поверхности катания головки рельса – это *случайная величина!!!*

Далее нужно отметить, что твердость - это непрерывная случайная величина, поскольку (если, например, рельсы отвечают требованиям первого класса) она может принимать любые значения из конечного интервала ($HV=341 \dots 388$, см. пункт 1.4 ГОСТ 18267-82).

После этого можно выдвинуть гипотезу (предположение), что такая случайная величина, как твердость на поверхности катания головки рельса, не должна противоречить нормальному закону распределения. Согласно центральной предельной теореме математической статистики, данную гипотезу скорее всего можно будет принять в качестве рабочей, поскольку опытные данные в примере 3.1 получены при измерении твердости в различных точках по длине одного и того же рельса. Следовательно, наиболее существенные факторы, которые определяют механические свойства данного металла на всех стадиях технологического процесса (получение металла, прокатка, термическая обработка), зафиксированы на одних и тех же уровнях. Кроме того, отклик (твердость металла) становится случайной величиной только в результате влияния малозначимых неконтролируемых факторов, число которых на различных этапах металлургического цикла, по всей видимости, стремится к бесконечности.

Итак, в качестве ответа на вопрос примера 3.1 мы можем сказать, что твердость на поверхности катания головки рельса – это непрерывная случайная величина, функцию распределения которой скорее всего можно записать в виде

$$F (HB) = \frac{1}{\sqrt{2 \pi \sigma_{HB}^2}} \int_{-\infty}^x e^{-\frac{(x - M_{HB})^2}{2 \sigma_{HB}^2}} dx .$$

Теперь, казалось бы, только осталось подсчитать по (2.15) математическое ожидание M_{HB} и по (2.17) – дисперсию σ_{HB}^2 , т.е. два параметра этой случайной величины, и у нас появится возможность определять вероятность того, что твердость на поверхности катания головки рельса принадлежит к некоторому заданному интервалу (например, $HB = 341 \dots 388$).

Однако на данном этапе мы попадаем в какой -то замкнутый круг: ведь для того, чтобы записать функцию нормального распределения, необходимо определить математическое ожидание и дисперсию; для вычисления этих двух параметров нужно знать плотность распределения (см. (2.15) и (2.17)), а плотность распределения - это первая производная от функции распределения (см.

(2.7)), т.е. в итоге, для того, чтобы найти функцию распределения, нужно знать функцию распределения.

Выход из подобного замкнутого круга может быть найден только лишь после того, как будет определена причина, по которой мы в него попадаем.

Итак, нам необходима функция распределения, причем для начала пусть хотя бы одно из ее значений, например $F(341)$. По определению (2.3) – это вероятность того, что случайная величина HV принимает значение не более 341. В свою очередь вероятность данного события $F(341) = P(HV \leq 341)$ есть предел частоты реализации события $HV \leq 341$ (отношение числа наблюдений, в которых твердость на поверхности катания головки рельса оказалась не более 341, к общему количеству наблюдений) при неограниченном числе повторений одного и того же комплекса условий.

А вот неограниченным числом повторений (генеральной совокупностью) в условиях примера 3.1 мы как раз и не располагаем, поскольку имеется только лишь три участка (сечения) рельса (три наблюдаемых единицы), в которых определена твердость на поверхности катания головки (три результата наблюдения).

Наблюдаемая единица – действительный или условный предмет, над которым проводят серию наблюдений.

Результат наблюдения – характеристика свойств единицы, полученная опытным путем.

Генеральная совокупность – множество всех рассматриваемых единиц.

Другими словами, генеральная совокупность - это такое воображаемое, в пределе бесконечно большое число предметов, над которыми можно провести наблюдения при неограниченном числе повторений одного и того же комплекса условий.

В примере 3.1 под генеральной совокупностью можно понимать, допустим, все участки одного и того же рельса, в которых в принципе можно было бы измерить твердость, либо вообще все рельсы Р65, которые когда-либо изготавливались или еще будут производиться по ГОСТ 18267-82.

В распоряжении исследователя, конечно же, никогда нет генеральной совокупности, и он может изучать только ее часть – выборку, причем всегда ограниченного объема.

Выборка – любое конечное подмножество генеральной совокупности, предназначенное для непосредственных исследований.

Объем – количество единиц в выборке.

По выборке невозможно однозначно определить ни функцию распределения, ни плотность распределения, ни параметры распределения (например, математическое ожидание или дисперсию) случайной величины, поскольку для этого потребуется неограниченное (бесконечно большое) количество результатов наблюдений, т.е. необходимо исследовать всю генеральную совокупность.

Следовательно, имея конечное подмножество генеральной совокупности (выборку), мы должны либо вообще отказаться от поиска распределения исследуемой случайной величины, либо удовлетвориться лишь некоторыми приближенными значениями неизвестных параметров ее распределения, т.е. провести оценивание случайной величины.

Оценивание – определение приближенного значения неизвестного параметра генеральной совокупности по результатам наблюдений.

Идея оценивания должна быть вполне понятна из соображений обычной житейской практики. Ведь для того, чтобы, например, купить пару килограмм яблок, у нас никогда не возникает желание съесть все имеющиеся у данного продавца фрукты (изучить всю генеральную совокупность), мы пробуем дольку только лишь одного яблока (исследуем выборку), определяем ее вкус (оцениваем) и принимаем решение, стоит нам или нет покупать именно эти яблоки.

Исходными данными при оценивании, как и при проверке любых предположений (статистических гипотез), касающихся неизвестного распределения случайной величины, конечно же, могут быть лишь только те результаты наблюдений, которые были получены в ходе проведения опытов (на выборке ограниченного объема). Причем предварительная обработка экспериментальных данных обычно начинается с подсчета тех или иных функций от результатов наблюдений (статистик).

Статистика – функция результатов наблюдений, используемая для оценки параметров распределения и (или) для проверки статистических гипотез.

По выборке невозможно найти параметры распределения случайной величины (поскольку для этого требуется бесконечное количество результатов наблюдений – изучение всей генеральной совокупности), поэтому, имея в своем распоряжении всегда ограниченный объем экспериментальных данных, исследователю остается довольствоваться только лишь получением некоторых оценок.

Оценка – статистика, являющаяся основой для оценивания неизвестного параметра распределения.

Для одного и того же параметра распределения может быть предложено несколько оценок.

В примере 3.1 рассматривалось четыре различных оценки для такого параметра распределения твердости, как математическое ожидание данной случайной величины (выборочное среднее арифметическое, выборочное среднее геометрическое, середина размаха и средний член вариационного ряда).

Поэтому при оценивании всегда возникает проблема выбора наилучшей оценки из всех возможных оценок данного параметра.

Причем, когда формулируются те или иные требования, по которым оценку целесообразно считать наилучшей, прежде всего учитывается тот факт, что любая оценка – это также случайная величина.

Ведь если бы в условиях примера 3.1 было бы найдено, допустим, выборочное среднее арифметическое твердости на поверхности катания головки какого-либо другого рельса, то, конечно же, совершенно не обязательно, что оно опять оказалось бы равно именно 362,00 единицам по Бринеллю.

Из тех соображений, что любая оценка Θ^* какого-либо параметра распределения Θ случайной величины тоже есть случайная величина, к оценкам предъявляются требования состоятельности, несмещенности и эффективности.

Состоятельная оценка – оценка, сходящаяся по вероятности к значению оцениваемого параметра при безграничном возрастании объема выборки.

$$\lim_{n \rightarrow \infty} \Theta^*(n) = \Theta, \quad (3.1)$$

где Θ - оцениваемый параметр; Θ^* - оценка; n - объем выборки.

Иными словами, для состоятельной оценки Θ^* отклонение ее от Θ на малую величину ε и более становится маловероятным при большом объеме выборки.

Вполне естественно, что исследователей в первую очередь интересуют те оценки, которые хотя бы в пределе (при проведении бесконечно большого количества наблюдений) давали им возможность определить интересующий их параметр распределения, т.е. чтобы оценки прежде всего были состоятельными. Однако следует отметить, что на практике приходится оценивать неизвестные параметры и при малых объемах выборки.

Естественным является требование, при выполнении которого оценка не дает систематической погрешности в сторону завышения (или занижения) истинного значения параметра Θ .

Несмещенная оценка – оценка, математическое ожидание которой равно значению оцениваемого параметра:

$$M(\Theta^*) = \Theta. \quad (3.2)$$

Удовлетворение требованию несмещенности позволяет устранить систематическую погрешность оценки параметра, которая зависит от объема выборки n и в случае состоятельности оценки стремится к нулю при $n \rightarrow \infty$.

Эффективная оценка – несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра.

$$M\{(\Theta^* - \Theta)^2\} = \min \quad (3.3)$$

или

$$M\{(\Theta^* - \Theta)^2\} \leq M\{(\Theta_i^* - \Theta)^2\} \quad (3.4)$$

где Θ_i^* – любая другая оценка.

Иными словами, дисперсия эффективной оценки параметра в некотором классе является минимальной среди дисперсий всех оценок из рассматриваемого класса несмещенных оценок.

Из всех состоятельных и несмещенных оценок следует предпочесть такую, которая оказывается наиболее близкой к оцениваемому параметру (эффективной), однако используемые в математической статистике оценки не всегда одновременно удовлетворяют всем трем перечисленным выше требованиям.

После того как исследователь выбрал и подсчитал состоятельную, несмещенную и эффективную оценки интересующего его параметра распределения исследуемой случайной величины, первое и наиболее простое, что он может сделать, так это принять значение оценки как неизвестное значение параметра распределения, т.е. выполнить точечное оценивание.

Точечное оценивание – способ оценивания, заключающийся в том, что значение оценки принимают как неизвестное значение параметра распределения.

Рассмотрим некоторые точечные оценки основных параметров распределения для непрерывной случайной величины, не противоречащей нормальному закону распределения.

Выборочное среднее арифметическое \bar{x} – сумма значений рассматриваемой величины, полученных по результатам испытания выборки, деленная на ее объем.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad i = 1, 2, \dots, n, \quad (3.5)$$

где n – объем выборки; x_i – результат измерения i -й единицы.

В математической статистике доказано, что выборочное среднее арифметическое является наилучшей (состоятельной, несмещенной и эффективной) оценкой математического ожидания случайной величины, подчиняющейся нормальному закону распределения.

В примере 3.1, даже если предположить, что твердость на поверхности катания головки рельса не противоречит нормальному закону распределения, из четырех полученных оценок предпочтение следует отдать значению $HВ_1 = (351 + 370 + 365)/3 = 362,00$ (выборочному среднему арифметическому) как наилучшей оценке для математического ожидания данной случайной вели-

чины. Три другие рассмотренные в этом примере оценки также являются состоятельными для математического ожидания. Однако среднее геометрическое $HB_2 = \sqrt[3]{351 \cdot 370 \cdot 365} = 361,91$ – это смещенная оценка (она будет наилучшей только тогда, когда случайная величина подчиняется так называемому логарифмически нормальному распределению, т.е. когда закону Гаусса подчиняется не сама случайная величина, а ее логарифм). Середина размаха $HB_3 = (351 + 370) / 2 = 360,50$ и средний член вариационного ряда $HB_4 = 365,00$ – это хотя и несмещенные оценки для математического ожидания, но их эффективность, как показано в математической статистике, меньше, чем у выборочного среднего арифметического (меньше единицы).

Выборочная дисперсия S_x^2 или $\overset{\circ}{S}_x^2$ – сумма квадратов отклонений выборочных результатов наблюдений от их выборочного среднего арифметического в выборке, деленная на $n-1$ или на n .

$$S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (3.6)$$

или

$$\overset{\circ}{S}_x^2 = \frac{\sum_{i=1}^n (x_i - M_x)^2}{n}. \quad (3.7)$$

Оценки S_x^2 и $\overset{\circ}{S}_x^2$ являются состоятельными, несмещенными и, в случае нормального распределения, асимптотически эффективными оценками дисперсии σ^2 .

Для практических расчетов выражение (3.6) можно преобразовать к виду

$$S_x^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]. \quad (3.8)$$

В условиях примера 3.1 выборочная дисперсия твердости на поверхности катания головки рельса равна

$$S_{HB}^2 = \frac{1}{3-1} \left[(351^2 + 370^2 + 365^2) - \frac{1}{3} (351 + 370 + 365)^2 \right] = 97,00.$$

Выборочное среднее квадратичное отклонение S_x или $\overset{\circ}{S}_x$ – положительный квадратный корень из выборочной дисперсии:

$$\overset{\circ}{S}_x = +\sqrt{S_x^2} \quad \text{или} \quad (3.9)$$

$$S_x = +\sqrt{S_x^2}. \quad (3.10)$$

В примере 3.1 $S_x = +\sqrt{97} = 9,85$.

Зная выборочное среднее арифметическое \bar{x} и выборочное среднее квадратичное отклонение S_x , можно подсчитать меру относительной изменчивости случайной величины – выборочный коэффициент вариации ν - по формуле

$$\nu = \frac{S_x}{\bar{x}}, \quad (3.11)$$

или, в процентах,

$$\nu = \frac{S_x}{\bar{x}} \cdot 100\%. \quad (3.12)$$

Для примера 3.1 выборочный коэффициент вариации твердости равен $\nu = 9,85/362 = 0,027$, или 2,7%.

Через выборочное среднее арифметическое \bar{x} и выборочное среднее квадратическое отклонение S_x могут быть сделаны точечные оценки для любых значений функции распределения, а также для вероятности попадания случайной величины в любой из заданных интервалов.

Так, для какого - либо значения функции нормального распределения, поскольку

$$F(x) = P(X \leq x) = P\left(\frac{X - M_x}{\sigma} \leq \frac{x - M_x}{\sigma}\right) = P\left(Z \leq \frac{x - M_x}{\sigma}\right) = \Phi\left(\frac{x - M_x}{\sigma}\right), \quad (3.13)$$

в качестве точечной оценки $F(x)$ можно использовать

$$\hat{F}(x) = \Phi\left(\frac{x - \bar{x}}{S_x}\right). \quad (3.14)$$

Точечную оценку вероятности попадания случайной величины X с нормальным законом распределения в любой из заданных интервалов (x_1, x_2) можно найти по формуле

$$\hat{P}(x_1 < X \leq x_2) = \Phi\left(\frac{x_2 - \bar{x}}{S_x}\right) - \Phi\left(\frac{x_1 - \bar{x}}{S_x}\right). \quad (3.15)$$

В соответствии с (2.32) точечная оценка квантили x_p порядка p для нормального распределения равна

$$\hat{x}_p = \bar{x} + z_p S_x. \quad (3.16)$$

В примере 3.1 предположим, что получено только два значения твердости на поверхности катания головки рельса (на обоих концах на расстоянии не более 1 м от торцов): 351 и 370, а третье испытание (в средней части) еще не проводилось.

Оценим при этих условиях вероятность того, что после измерения твердости в средней части рельса ее значение окажется ниже, чем 341, т.е. вероятность того, что в результате третьего испытания рельс попадет во второй класс (для которого твердость на поверхности катания головки может лежать в диапазоне 311...341) или его придется подвергнуть повторной однократной термической обработке (закалке и отпуску).

Кроме того, оценим вероятность того, что после определения твердости в средней части рельса он будет по-прежнему удовлетворять требованиям первого класса по пункту 1.4 ГОСТ 18267-82 ($HV = 341 \dots 388$).

Если предположить, что твердость на поверхности катания головки рельса не противоречит нормальному закону распределения, то наилучшими точечными оценками для математического ожидания и дисперсии этой случайной величины в соответствии с (3.5) и (3.8) будут значения

$$\bar{x}_{HV} = \frac{1}{2}(351 + 370) = 360,5,$$

$$S_{HB}^2 = \frac{1}{2-1} \left[(351^2 + 370^2) - \frac{1}{2} (351 + 370)^2 \right] = 180,5,$$

а по (3.10) выборочное среднее квадратичное отклонение составит $S_x = +\sqrt{180,5} = 13,435$.

Тогда по (3.14) получаем, что

$$F^*(341) = P^*(HB \leq 341) = \Phi\left(\frac{341 - 360,5}{13,435}\right) = \Phi(-1,45).$$

Поскольку согласно (2.29) $\Phi(-1,45) = 1 - \Phi(1,45)$, то по таблицам для функции Лапласа (см. прил. П1) находим $\Phi(1,45) = 0,92647$, и, следовательно, $F^*(341) = \Phi(-1,45) = 1 - 0,92647 \approx 0,07$.

В электронных таблицах Microsoft Excel для подобных расчетов можно использовать функцию **НОРМСТРАСП**.

$$\text{НОРМСТРАСП}((341 - \text{СРЗНАЧ}(351;370)) / \text{СТАНДОТКЛОН}(351;370)) = 0,07333,$$

где **СРЗНАЧ(351;370)** и **СТАНДОТКЛОН(351;370)** – статистические функции для вычисления соответственно выборочного среднего арифметического значения и выборочного среднего квадратичного отклонения.

Точно такое же значение может быть получено через функцию **НОРМРАСП**.

$$\text{НОРМАСП}(341; \text{СРЗНАЧ}(351;370); \text{СТАНДОТКЛОН}(351;370); \text{ИСТИНА}) = 0,0733.$$

Итак, точечной оценкой (полученной по двум выборочным значениям 351 и 370) функции распределения твердости *HB* на поверхности катания головки рельса от значения 341 является величина $F^*(341) = P^*(HB \leq 341) = 0,07$. Другими словами, точечная оценка вероятности того, что при испытании твердости в средней части рельса ее значение окажется меньше 341, равна 0,07. Или, если проведено два испытания на обоих концах рельса и получены значения 351 и 370, то после определения твердости в средней части, возможно, только семь рельсов из ста придется перевести во второй класс или подвергнуть повторной однократной термической обработке.

Для оценки вероятности того, что после определения твердости в средней части рельс по-прежнему будет удовлетворять требованиям первого класса, воспользуемся соотношением (3.15) и получим

$$P^*(341 < HB \leq 388) = \Phi\left(\frac{388 - 360,5}{13,435}\right) - \Phi\left(\frac{341 - 360,5}{13,435}\right) = \Phi(2,05) - \Phi(-1,45).$$

Значение $\Phi(-1,45) \approx 0,07$ нами было уже найдено, а $\Phi(2,05) \approx 0,98$ (по таблицам [11], в табл. П.1 или в Microsoft Excel **НОРМСТРАСП(2,05) = 0,979818**).

Следовательно, $P^*(341 < HB \leq 388) \approx 0,98 - 0,07 = 0,91$, т.е. 91% всех рельсов, после измерения твердости в средней части, будут по-прежнему отвечать требованиям пункта 1.4 ГОСТ 18267-82 ($HB = 341 \dots 388$), если на их концах уже были получены значения 351 и 370.

Добавим, что значения 341 и 388 являются оценками квантилей порядка соответственно 0,07 и 0,98, т.е. $HB_{0,07}^* = 341$, а $HB_{0,98}^* = 388$, и если, допустим, необходимо оценить квантиль порядка 0,99, то по формуле (3.16) можно получить следующее значение:

$$HB_{0,99}^* = 360,5 + z_{0,99} \cdot 13,435 = 360,5 + 2,326 \cdot 13,435 = 391,80,$$

где $z_{0,99}$ – квантиль нормированного нормального распределения порядка 0,99 – можно найти по таблицам [11], в табл. П.2 или в Microsoft Excel с использованием функции **НОРМСТОБР(0,99) = 2,326342**, а также **НОРМОБР(0,99;0;1) = 2,326342**.

Следовательно, если на обоих концах рельса получены значения 351 и 370, то скорее всего только в одном случае из ста твердость на поверхности катания головки в средней части может оказаться больше 391,8.

Однако все последние приведенные в примере 3.1 выводы и заключения относительно оценок различных вероятностей не следует понимать в буквальном смысле слова. Так, если бы удалось собрать данные по твердости в средней части на ста рельсах, у которых значения этого показателя качества по концам составляли бы ровно 351 и 370, то, конечно же, совершенно не обязательно, что именно только на одном рельсе из ста твердость оказалась бы

больше, чем 391,8. Такое событие вполне могло бы быть отмечено и на двух, и на трех и т.д. рельсах либо вообще ни разу не встретиться.

Дело здесь заключается в том, что, во-первых, даже если бы нам удалось найти самую теоретическую вероятность какого-либо события (изучить всю генеральную совокупность), а не ее оценку (полученную по выборке ограниченного объема), то и в этом случае фактическая частота реализации этого события вполне могла бы отличаться (хотя и не очень сильно) от соответствующей ей теоретической вероятности. Так, например, если сто раз подбросить "идеальную" монету, то совершенно не обязательно, что ровно в 50 случаях выпадет "орел", а в остальных 50 - "решка". Хотя то, что во всех 100 случаях выпадет "орел" и ни разу - "решка", мы вряд ли увидим (если тот, кто подбрасывает монету, не факир или фокусник, то вероятность подобного события равна $(0,5)^{100} = 8 \cdot 10^{-31}$).

И, во-вторых, если в нашем распоряжении имеются только лишь какая-либо точечная оценка, то вообще совершенно невозможно сказать, насколько близко она располагается относительно оцениваемого ею параметра. Так, например, если вероятность того, что при получении твердости на концах рельса 351 и 370 он и после измерения этой величины в средней части будет отвечать пункту 1.4 ГОСТ 18267-82 оценивается значением 0,91, то на самом деле (для всей генеральной совокупности, т.е. для всех рельсов Р65, выпускаемых по ГОСТ 18267-82) эта вероятность может быть равна и 0,85, и 0,95 и т.д.

По значению точечной оценки не представляется возможным определить, хотя бы в каком диапазоне находится оцениваемый ею параметр. Этот существенный недостаток точечного оценивания может быть компенсирован оцениванием с помощью так называемого доверительного интервала.

3.2. Оценивание с помощью доверительного интервала

В отличие от точечной оценки, интервальная оценка позволяет получить вероятностную характеристику точности оценивания неизвестного параметра.

Идея оценивания с помощью доверительного интервала заключается в том, чтобы в окрестности точечной оценки попытаться построить такой интер-

вал (доверительный интервал), который с некоторой, отличной от нуля, вероятностью (доверительной вероятностью) накрыл бы оцениваемый параметр распределения.

Доверительный интервал – интервал, который с заданной вероятностью накроет неизвестное значение оцениваемого параметра распределения.

Доверительная вероятность – вероятность того, что доверительный интервал накроет действительное значение параметра, оцениваемого по выборочным данным.

Оценивание с помощью доверительного интервала – способ оценки, при котором с заданной доверительной вероятностью устанавливают границы доверительного интервала.

Предположим, что для оценки параметра Θ удалось найти две функции $\Theta_1^*(x_1, x_2, \dots, x_n)$ и $\Theta_2^*(x_1, x_2, \dots, x_n)$, такие, что при всех (x_1, x_2, \dots, x_n) и при любых значениях Θ выполняется условие

$$\left. \begin{aligned} \Theta_1^* < \Theta_2^*; \\ P\left\{\Theta_1^*(x_1, x_2, \dots, x_n) \leq \Theta \leq \Theta_2^*(x_1, x_2, \dots, x_n)\right\} = 1 - \alpha. \end{aligned} \right\} \quad (3.17)$$

Это означает, что действительное значение параметра Θ находится в интервале значений $(\Theta_1^*; \Theta_2^*)$ с вероятностью P .

Интервал $(\Theta_1^*; \Theta_2^*)$ как раз и называют доверительным интервалом для неизвестного параметра Θ , а соответствующую ему вероятность $P\{\Theta_1^* \leq \Theta \leq \Theta_2^*\}$ – доверительной вероятностью (или надежностью) $P=1-\alpha$, где α – уровень значимости. Если, к примеру, $\alpha = 0,05$, то строится доверительный интервал с доверительной вероятностью 0,95 (или 95-процентный доверительный интервал).

Часто доверительный интервал находится как интервал, симметричный относительно точечной оценки параметра. Для симметричного доверительного интервала его ширина 2δ определяется условием

$$P\left\{|\Theta - \bar{\Theta}^*| \leq \delta\right\} = 1 - \alpha, \quad (3.18)$$

где $\bar{\Theta}^*$ – точечная оценка параметра Θ .

При фиксированном значении α (вероятности того, что доверительный интервал не покроет действительного значения параметра) чем меньше δ , тем точнее оценивается Θ .

Вероятностное утверждение $P\{\Theta_1^* \leq \Theta \leq \Theta_2^*\}$ не следует понимать таким образом, что параметр Θ есть случайная величина, которая с вероятностью P попадет в интервал между Θ_1^* и Θ_2^* .

Любой параметр распределения Θ (в отличие от его оценок) – это детерминированная величина, неизвестная нам, но имеющая строго определенное, фиксированное значение (которое, по крайней мере, теоретически, может быть найдено при исследовании всей генеральной совокупности). Границы Θ_1^* и Θ_2^* (как некоторые функции от результатов наблюдений) есть случайные величины. Поэтому утверждение $P\{\Theta_1^* \leq \Theta \leq \Theta_2^*\} = P$ означает, что для данного доверительного интервала $(\Theta_1^*; \Theta_2^*)$ вероятность содержать значение Θ равна P .

Рассмотрение способов получения интервальных оценок для основных параметров распределения начнем с построения доверительного интервала для математического ожидания, так как именно такие задачи наиболее часто встречаются в инженерной практике.

3.2.1. Построение доверительного интервала для математического ожидания

Как уже было отмечено, наилучшей (состоятельной, несмещенной и эффективной) точечной оценкой математического ожидания случайной величины X с нормальным законом распределения является ее выборочное среднее арифметическое \bar{x} . Поэтому за основу построения доверительного интервала для математического ожидания обычно выбирается именно эта точечная оценка данного параметра. Задача получения интервальной оценки в этом случае заключается в поиске границ $(\bar{x} - \delta; \bar{x} + \delta)$ такого интервала, который с заданной доверительной вероятностью P_{Mx} покроет действительное значение математического ожидания M_x (рис.3.1).

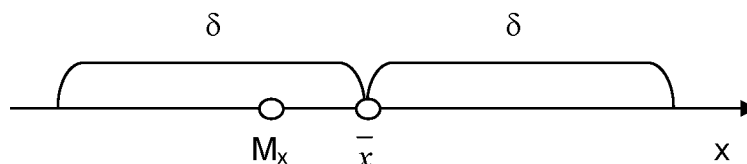


Рис.3.1. Построение доверительного интервала для математического ожидания

При построении любой интервальной оценки, в том числе и для математического ожидания, необходимо знать распределение той точечной оценки (случайной величины), которая берется за основу для построения доверительного интервала.

В математической статистике доказано, что выборочное среднее арифметическое \bar{x} из n независимых результатов наблюдений случайной величины, распределенной нормально с параметрами M_x и σ_x^2 , также подчиняется нормальному закону распределения с параметрами:

$$M(\bar{x}) = M_x, \quad (3.19)$$

$$\sigma^2(\bar{x}) = \sigma_x^2 / n. \quad (3.20)$$

Подтвердить справедливость равенства (3.19) можно хотя бы тем, что выборочное среднее арифметическое - это несмещенная оценка математического ожидания, следовательно, по определению (см. (3.2)), математическое ожидание этой оценки (выборочного среднего арифметического) равно значению оцениваемого параметра (математическому ожиданию).

Соотношение (3.20) не должно, интуитивно, вызывать ни каких серьезных возражений: ведь если подсчитать выборочное среднее арифметическое по нескольким выборкам одного и того же объема, а затем найти дисперсию полученных значений, то вероятнее всего предположить, что разброс (дисперсия) выборочных средних арифметических будет меньше, чем разброс (дисперсия) самих опытных данных.

Прокомментируем это положение следующим иллюстративным числовым материалом (в продолжение примера 3.1). На каждом двадцатом по ходу технологического процесса рельсе Р65 (по ГОСТ 18267-82) получены следующие значения твердости на поверхности катания головки:

первый рельс	– 351, 370, 365 ($\bar{x}_{HB} = 362$, $S_{HB}^2 = 97$);
двадцать первый рельс	– 375, 369, 345 ($\bar{x}_{HB} = 363$, $S_{HB}^2 = 252$);
сорок первый рельс	– 348, 363, 369 ($\bar{x}_{HB} = 360$, $S_{HB}^2 = 117$).

Если теперь по (3.8) оценить дисперсию такой случайной величины, как \bar{HB} , то получим

$$S_{\bar{HB}}^2 = \frac{1}{3-1} \left[(362^2 + 363^2 + 360^2) - \frac{1}{3} (362 + 363 + 360)^2 \right] = 2,33.$$

Как видно из этого числового примера, выборочная дисперсия средних арифметических – 2,33 по трем выборкам (объемом 3) почти на порядок меньше тех выборочных дисперсий (97, 252 и 117), которые имеют сами опытные данные.

Для более строгого обоснования соотношения (3.20) напомним, что если случайная величина $Y = X_1 \pm X_2$ – является суммой или разностью двух независимых случайных величин X_1 и X_2 , то справедливо равенство

$$\sigma_y^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2. \quad (3.21)$$

Кроме того, дисперсия произведения случайной переменной X и постоянной величины (константы) C равна

$$\sigma^2(C \cdot x) = C^2 \cdot \sigma_x^2. \quad (3.22)$$

Закон сложения дисперсий справедлив при любом числе слагаемых. Учитывая, что $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и σ_x^2 – дисперсия случайной величины X , а также соотношения (3.21) и (3.22), получаем:

$$\sigma^2(\bar{x}) = \sigma^2\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \left(\frac{1}{n}\right)^2 \sigma^2\left(\sum_{i=1}^n x_i\right) = \frac{n \cdot \sigma_x^2}{n^2} = \frac{\sigma_x^2}{n},$$

что и требовалось доказать, причем $\sigma(\bar{x}) = \sigma_x \cdot \frac{1}{\sqrt{n}}$.

Если заранее известна дисперсия σ_x^2 , то доверительный интервал для математического ожидания M_x рассчитывается достаточно просто. Его границы можно найти, например, следующим образом.

Поскольку случайная величина \bar{X} подчиняется нормальному закону распределения с параметрами $M(\bar{x}) = M_x$ и $\sigma^2(\bar{x}) = \sigma_x^2/n$, то соответствующая ей приведенная случайная величина

$$Z = \frac{\bar{X} - M(\bar{x})}{\sigma(\bar{x})} = \frac{\bar{X} - M_x}{\sigma_x / \sqrt{n}}, \quad (3.23)$$

имеет нормированный стандартный нормальный закон распределения [см.(2.27)].

Квантиль \bar{x}_p порядка P такой случайной величины, как \bar{X} , определяется аналогично (2.32а) и с учетом соотношений (3.19) и (3.20) равна:

$$\bar{x}_p = M(\bar{x}) + z_p \sigma(\bar{x}) = M_x + z_p \frac{\sigma_x}{\sqrt{n}}.$$

Далее, в соответствии с (2.20)

$$P(\bar{x}_{P_1} < \bar{x} < \bar{x}_{P_2}) = P\left(M_x + z_{P_1} \frac{\sigma_x}{\sqrt{n}} < \bar{x} < M_x + z_{P_2} \frac{\sigma_x}{\sqrt{n}}\right) = P_2 - P_1.$$

Если в последнем соотношении неравенство, стоящее под знаком вероятности, разрешить относительно M_x , то получим

$$P\left(\bar{x} - z_{P_2} \frac{\sigma_x}{\sqrt{n}} < M_x < \bar{x} - z_{P_1} \frac{\sigma_x}{\sqrt{n}}\right) = P_2 - P_1. \quad (3.24)$$

Если $\bar{x} < M_x + z_{P_2} \frac{\sigma_x}{\sqrt{n}}$, то $-M_x < -\bar{x} + z_{P_2} \frac{\sigma_x}{\sqrt{n}}$ и, следовательно, $M_x > \bar{x} - z_{P_2} \frac{\sigma_x}{\sqrt{n}}$,

и, аналогично, если $\bar{x} > M_x + z_{P_1} \frac{\sigma_x}{\sqrt{n}}$, то $-M_x > -\bar{x} + z_{P_1} \frac{\sigma_x}{\sqrt{n}}$,

и, следовательно, $M_x < \bar{x} - z_{P_1} \frac{\sigma_x}{\sqrt{n}}$.

Таким образом, вероятность того, что выполняется неравенство

$$\bar{x} - z_{P_2} \frac{\sigma_x}{\sqrt{n}} < M_x < \bar{x} - z_{P_1} \frac{\sigma_x}{\sqrt{n}}, \quad (3.25)$$

будет $P = P_2 - P_1 = 1 - \alpha$.

Если для примера принять $P_1 = 0,025$ и $P_2 = 0,975$ ($P = 0,975 - 0,025 = 0,95$; $\alpha = 0,05$), то, поскольку (см. (2.32)) $z_{0,025} = z_{1-0,975} = -z_{0,975}$, а $z_{0,975} = 1,96$ (по таблицам [11], табл. П.2 или используя **НОРМСТОБР(0,975) = 1,959961**), получим

$$P\left(\bar{x} - 1,96 \frac{\sigma_x}{\sqrt{n}} < M_x < \bar{x} + 1,96 \frac{\sigma_x}{\sqrt{n}}\right) = 0,95, \quad (3.26)$$

т.е. при многократном извлечении выборок (объемом n каждая) из нормально распределенной генеральной совокупности (с параметрами M_x и σ_x^2) можно построить последовательность соответствующих данным выборкам интервалов (3.26), причем примерно 95% этих интервалов будут включать в себя (накрывать) истинное значение математического ожидания M_x .

При построении доверительного интервала для математического ожидания обычно принимают $P_1 = \alpha/2$ и $P_2 = 1 - \alpha/2$, т.е. рассматривают симметричные границы относительно выборочного среднего арифметического. В инженерных приложениях для значений α обычно выбирают $\alpha = 0,1$ или $\alpha = 0,05$, реже $\alpha = 0,01$, т.е. строят такие доверительные интервалы, которые в 90 или 95% (реже 99%) случаев накрывают математическое ожидание.

С учетом соотношения (2.32) $z_{\alpha/2} = -z_{1-\alpha/2}$, по (3.25) получаем, что вероятность выполнения неравенства

$$\bar{x} - z_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}} < M_x < \bar{x} + z_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}} \quad (3.27)$$

равна $P = 1 - \alpha/2 - \alpha/2 = 1 - \alpha$.

Следовательно, интервал (3.27) является доверительным интервалом для математического ожидания M_x случайной величины с нормальным законом распределения, построенным с доверительной вероятностью $P = 1 - \alpha$. Границы

этого интервала равны $\bar{x} - z_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}}$ и $\bar{x} + z_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}}$, а половина его ширины

(см. рис.3.1)
$$\delta = z_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}}.$$

Пример 3.2. При проектировании системы управления базой данных было проведено исследование характеристик файлов ряда действующих и разрабатываемых информационных систем. В процессе исследования рассмотрены $n = 49$ файлов и получены следующие данные: средняя величина файла $\bar{x} = 55$ кбайт, $\sigma(x) = 11$. Необходимо определить доверительный интервал M_x ; объем выборки n , который необходимо выполнить, чтобы точность статистических выводов $\delta \leq 2$, и величину записи R на физическом уровне хранения данных, обеспечивающую размещение файлов с надежностью $P=0,95$.

Воспользовавшись соотношением (3.26), рассчитаем доверительный интервал:

$$55 - 1,96 \frac{11}{\sqrt{49}} < M_x < 55 + 1,96 \frac{11}{\sqrt{49}},$$

$$51,9 \leq M_x \leq 58,1.$$

Длина записи $R = 55 + 1,96 * 11 \approx 77$ кбайт.

Необходимый объем выборки для $\delta = 2$ составит

$$n = \left(1,96 \frac{11}{2} \right)^2 \approx 116$$

На практике, как правило, число измерений (например, отбора проб шихты, чугуна, стали и других материалов) конечно и не превышает 10... 30. При таком малом числе наблюдений фактическая дисперсия σ_x^2 неизвестна, поэтому при построении доверительного интервала для математического ожидания M_x используют выборочную дисперсию S_x^2 .

В этом случае приведенная случайная величина, аналогичная (3.23),

$$t = \frac{(\bar{x} - M_x)}{S_x / \sqrt{n}}, \quad (3.27a)$$

где S_x – выборочное среднее квадратичное отклонение, определяемое по формуле (3.10), имеет распределение, отличное от нормального. Функция распределения случайной величины t (3.27) имеет вид

$$F(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{\pi m} \cdot \Gamma\left(\frac{m}{2}\right)} \int_{-\infty}^t \left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}} dt, \quad (3.28)$$

где $\Gamma(y)$ – гамма-функция, являющаяся обобщением понятия факториала и обладающая рекуррентным свойством: $\Gamma(y + 1) = y\Gamma(y)$ (для целых чисел n справедливо $\Gamma(n + 1) = n!$ см. [1]); m – число степеней свободы, определяемое разностью между объемом выборки n и числом параметров, оцениваемых по выборке; в данном случае $m = n - 1$ (поскольку при определении t по (3.27) необходимо оценить один параметр S_x).

Число степеней свободы m – это понятие, которое учитывает в статистических ситуациях связи, ограничивающие свободу изменения случайных величин. Поэтому число степеней свободы вычисляется как разность между числом экспериментальных точек n и числом связей f , ограничивающих свободу изменения случайной величины. Так, при вычислении выборочной дисперсии по формуле (3.6) $S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$ наблюдается одна связь, определяемая

уровнем выборочного среднего $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, поэтому число степеней свободы

выборочной дисперсии будет равно $m = n - 1$, а, например, для выборочной

дисперсии, найденной из соотношения (3.7) $\tilde{S}_x^2 = \sum_{i=1}^n (x_i - M_x)^2 / n$, число степеней

свободы равно числу испытаний $m = n$, так как M_x определено независимым способом.

Понятие о степени свободы поясним еще на примере решения системы линейных алгебраических уравнений. Допустим, что мы имеем систему из n линейных алгебраических уравнений с n неизвестными x_1, x_2, \dots, x_n . Очевидно,

решение такой системы (при линейной независимости уравнений) будет единственным, т.е. такая система не будет иметь ни одной степени свободы. Но если для n неизвестных переменных мы имеем только одно уравнение, то для однозначного определения x_1, x_2, \dots, x_n должно быть наложено еще $m = n - 1$ условий (уравнений), т.е. число степеней свободы такой системы уравнений будет равно $n - 1$.

Наконец, если по выборке объемом n будут сделаны оценки ровно для n (линейно независимых) параметров распределения, то расчет $n + 1$ оценки не будет нести никакой дополнительной информации о распределении случайной величины (все n выборочных значений x_1, x_2, \dots, x_n будут однозначно определены через n оценок параметров), поскольку после оценки n параметров число степеней свободы $m = n - n$ уже окажется равным нулю.

Распределение (3.28), зависящее только от числа степеней свободы (однопараметрическое), называют распределением Стьюдента, или t -распределением.

Плотность распределения Стьюдента выражается формулой

$$f(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{\pi m} \cdot \Gamma\left(\frac{m}{2}\right)} \left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}}, \quad (3.29)$$

причем множители при $\left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}}$ в $f(t)$ выбраны так, чтобы площадь под любой кривой $f(t)$ равнялась единице.

Стьюдент – псевдоним У.С. Госсета (1876-1937) – химика, работавшего в одной из пивоваренных фирм Великобритании. Он самостоятельно разработал статистику малых выборок. Поскольку в современной технике чаще всего исследуются небольшие по объему выборки (менее 30), то работа Стьюдента имеет большое практическое значение.

На рис. 3.2 приведено распределение Стьюдента для различных значений m . При $n \rightarrow \infty$ (практически при $n \geq 30$) распределение Стьюдента переходит в стандартное нормальное распределение с единичной дисперсией.

Для случайной величины t (3.27), в соответствии с (2.20), можно записать, что

$$P(t_{P_1} < t \leq t_{P_2}) = P\left(t_{P_1} < \frac{(\bar{x} - M_x)}{S_x / \sqrt{n}} \leq t_{P_2}\right) = P_2 - P_1, \quad (3.30)$$

где t_{P_1} и t_{P_2} – значения квантилей случайной величины t порядка p_1 и p_2 соответственно.

Если в соотношении (3.30), аналогично (3.24), разрешить относительно M_x неравенство, стоящее под знаком вероятности, и при построении доверительного интервала для математического ожидания принять симметричные границы $P_1 = \alpha/2$ и $P_2 = 1 - \alpha/2$, то получим, что вероятность выполнения неравенства

$$\bar{x} - t_{\alpha,m} \frac{S_x}{\sqrt{n}} < M_x < \bar{x} + t_{\alpha,m} \frac{S_x}{\sqrt{n}} \quad (3.31)$$

равна $P = 1 - \alpha$, где $t_{\alpha,m}$ – так называемый коэффициент Стьюдента (значение квантили статистики t (3.27) порядка $P = 1 - \alpha/2$ для числа степеней свободы $m = n - 1$).

Следовательно, интервал (3.31) является доверительным интервалом для математического ожидания M_x случайной величины с нормальным законом распределения, построенным с доверительной вероятностью $P = 1 - \alpha$, при известном значении генеральной дисперсии σ_x^2 .

Значения $t_{\alpha,m}$ табулированы (см., например, [11] или табл. П.6), их можно определить также, воспользовавшись статистической функцией **СТЮДРАСПОБР** из электронных таблиц Microsoft Excel, причем при $m > 30$ $t_{\alpha,m} \approx Z_{1 - \alpha/2}$. Так, при $\alpha = 0,05$ и $m = 31$ **СТЮДРАСПОБР**(0,05;31) = 2,039515, а **НОРМСТОБР**(1-0,05/2) = 1,959961.

Если в примере 3.1 по трем ($n = 3$, $m = n - 1 = 3 - 1 = 2$) выборочным значениям 351, 370 и 365 (первый рельс $\bar{x}_{HB} = 362$; $S_{HB} = 9,85$) было бы необходимо при $\alpha = 0,05$ построить доверительный интервал для математического ожидания твердости на поверхности катания головки рельса, то, если предположить, что твердость не противоречит нормальному закону распределения, и поскольку $t_{0,05,2} \approx 4,3$ (**СТЮДРАСПОБР**(0,05;2) = 4,302656), он оказался бы равным

$$362 - 4,3 \frac{9,85}{\sqrt{3}} < M_{HB} < 362 + 4,3 \frac{9,85}{\sqrt{3}}, \text{ или } M_{HB} = 362 \pm 24,45.$$

Следовательно, интервал [337,55; 386,45] с вероятностью $1 - 0,05 = 0,95$ покрывает математическое ожидание твердости на поверхности катания головки рельса.

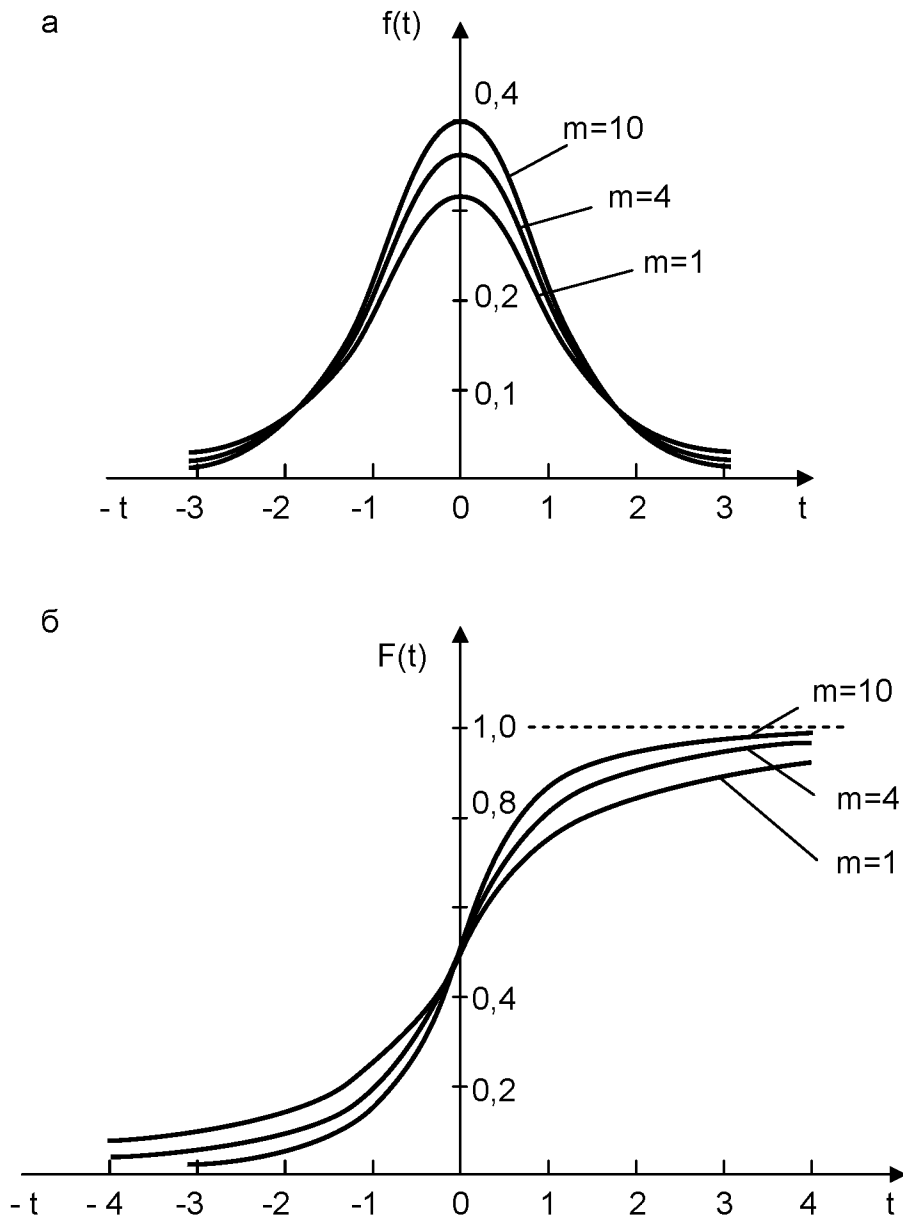


Рис. 3.2. Плотность (а) и функция (б) t -распределения Стьюдента

3.2.2. Построение доверительного интервала для дисперсии

При построении доверительного интервала для дисперсии используется случайная величина χ^2 (читается: "хи-квадрат"),

$$\chi^2 = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right)^2 = \frac{n-1}{\sigma_x^2} \cdot S_x^2, \quad (3.32)$$

которая имеет так называемое распределение Пирсона (по имени английского математика и биолога К.Пирсона), или χ^2 -распределение ("хи-квадрат-распределение").

Плотность распределения случайной величины χ^2 описывается уравнением

$$f(\chi^2) = \frac{1}{2^{m/2} \cdot \Gamma(m/2)} \cdot (\chi^2)^{\frac{m-2}{2}} \cdot e^{-\frac{1}{2}\chi^2}, \quad 0 \leq \chi^2 \leq \infty, \quad (3.33)$$

где $\Gamma(m/2)$ – гамма – функция; m – число степеней свободы (при построении доверительного интервала $m = n-1$).

На рис.3.3 приведены кривые $f(\chi^2)$ для различных значений m . Эти кривые асимметричны, причем асимметрия особенно резко выражена при малых значениях параметра m . Так, при $m = 1$ и $\chi^2 = 0$ кривая уходит в бесконечность, а при $m = 2$ и $\chi^2 = 0$ она достигает максимального значения, равного 0,5. При $m > 2$ кривые имеют максимум при $\chi^2_{\max} = m - 2$. При больших значениях m ($m > 30$) χ^2 -распределение переходит в нормальное со средним значением $\overline{f(\chi^2)} = \sqrt{2m - 1}$ и дисперсией $\sigma^2 = 1$.

Для построения доверительного интервала для дисперсии рассмотрим соотношение

$$P(\chi_{P_1}^2 < \chi^2 \leq \chi_{P_2}^2) = P_2 - P_1 \quad (3.34)$$

и с учетом (3.32) решим стоящее в скобках неравенство относительно σ_x^2 :

$$P\left(S_x^2 \frac{n-1}{\chi_{P_1}^2} < \sigma_x^2 \leq S_x^2 \frac{n-1}{\chi_{P_2}^2}\right) = P_2 - P_1, \quad (3.35)$$

$$\text{где } S_x^2 \frac{n-1}{\chi_{P_1}^2} < \sigma_x^2 < S_x^2 \frac{n-1}{\chi_{P_2}^2} - \quad (3.36)$$

есть доверительный интервал для дисперсии σ_x^2 с доверительной вероятностью $P = P_2 - P_1 = 1 - \alpha$.

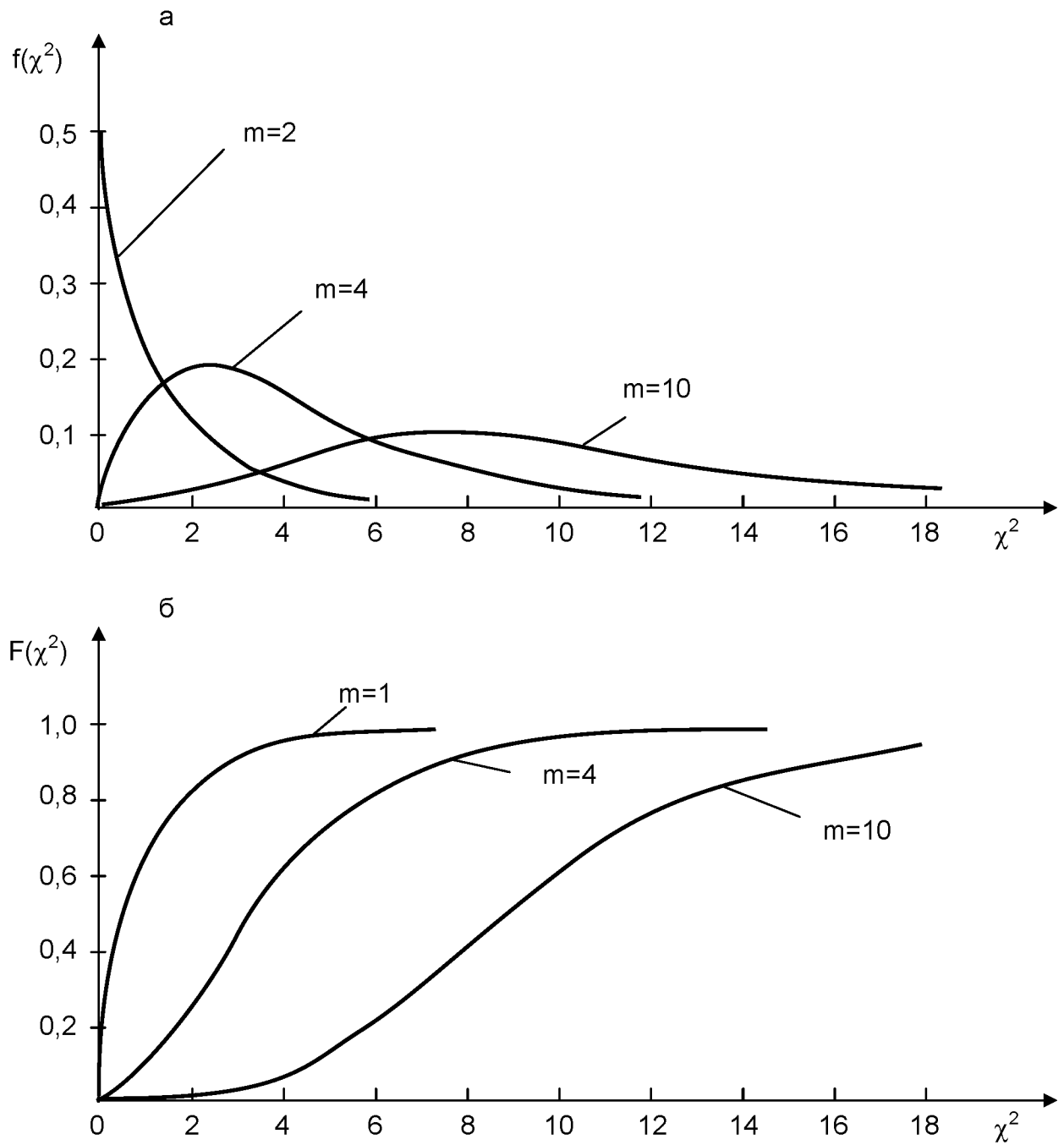


Рис.3.3. Плотность распределения (а) и функция распределения (б)

χ^2

Как и при построении доверительного интервала для математического ожидания в технических приложениях обычно принимают $P_1=\alpha/2$ и $P_2=1-\alpha/2$, а α выбирают равным 0,1 или 0,05, реже 0,01.

Квантили распределения Пирсона находят по таблицам (см. [11] или табл. П.3), а в Microsoft Excel для этого используется функция **ХИ2ОБР**.

Границы доверительного интервала для среднего квадратичного отклонения σ_x находятся путем извлечения квадратного корня из значений доверительных границ для дисперсии.

В примере 3.1 по трем выборочным значениям 351, 370 и 365, $S_{HB}^2 = 97$ при $\alpha = 0,05$ ($P_1=0,05/2=0,025$ и $P_2=1-0,05/2=0,975$; **ХИ2ОБР**(0,025;2) = 7,377779 и **ХИ2ОБР**(0,975;2) = 0,050636) доверительный интервал для дисперсии твердости составит

$$97 \frac{3-1}{7,38} < \sigma^2 < 97 \frac{3-1}{0,05}, \text{ или после вычислений } 26 < \sigma^2 < 3880, \text{ а довери-}$$

тельный интервал для среднего квадратичного отклонения будет равен $5 < \sigma < 62$.

3.2.3. Определение необходимого количества опытов при построении интервальной оценки для математического ожидания

Увеличение количества измерений (числа проб, образцов и т.п.), как видно из выражений (3.27) и (3.31) даже при неизменной их точности ($\sigma_x = \text{const}$), может увеличить доверительную вероятность P или сузить доверительный интервал $\pm\delta$ для определения действительного значения измеряемой величины (математического ожидания).

Необходимое количество измерений (образцов, проб и т.п.) n для достижения требуемой точности δ при заданной доверительной вероятности P можно определить заранее в том случае, когда известно действительное значение среднеквадратичного отклонения σ_x , а экспериментальные данные (измерения) подчиняются нормальному закону распределения.

Действительно, при этих допущениях число измерений можно определить из выражения (3.27)

$$n \geq \left(\frac{z_{1-\alpha/2} \cdot \sigma_x}{\delta} \right)^2 = z_{1-\alpha/2}^2 \cdot \left(\frac{\sigma_x}{\delta} \right)^2 = z_{1-\alpha/2}^2 \cdot (\varepsilon)^2, \quad (3.37)$$

где $\varepsilon = \sigma_x / \delta$.

Таким образом, число измерений n определяется требуемой доверительной вероятностью $P = 1 - \alpha$ и относительным (по отношению к среднеквадратичному отклонению) значением половины ширины доверительного интервала δ , т.е. требуемой точностью определения измеряемой величины. Так, при $P=0,95$, $z_{0,975} = 1,96$ и при $\delta = \sigma_x$ число измерений равно 4. При увеличении необходимой точности измерений в 2 раза, т.е. сужении доверительного интервала до величины $\delta = (1/2)\sigma_x$, необходимое число измерений составит 16. Нетрудно заметить, что необходимое число измерений с увеличением точности возрастает в квадратичной зависимости.

Как правило, действительное значение среднеквадратичной ошибки (σ_x) неизвестно, а имеется только ее оценка (S_x). В этом случае следует воспользоваться соотношением (3.31), т.е. критерием Стьюдента, и необходимое число измерений определять из соотношения

$$n \geq \frac{t_{\alpha,m}^2 \cdot S_x^2}{\delta^2} = t_{\alpha,m}^2 \cdot \left(\frac{S_x}{\delta} \right)^2 = t_{\alpha,m}^2 \cdot \varepsilon^2, \quad (3.38)$$

где $\varepsilon = S_x / \delta$.

При расчетах по этому уравнению следует иметь в виду, что значение критерия Стьюдента зависит не только от α , но и от числа степеней свободы m , последние же определяются числом измерений. В связи с этим уравнение (3.38) следует решать методом последовательных приближений. В качестве начального приближения можно задать, в частности, число измерений, рассчитанных по формуле (3.37). Так, если решить последнее уравнение методом последовательных приближений, то можно показать, что при $P=0,95$ ($\alpha=0,05$) для определения доверительного интервала с точностью $\delta = S_x$ требуется 7 измерений, а с точностью $\delta = 0,5S_x$ – 19. С повышением необходимой точности различие

в числе измерений, рассчитанных по соотношениям (3.37) и (3.38), уменьшается и, как показывают расчеты, при величине $\delta \leq 0,2S_x$ они практически совпадают.

В примере 3.1 доверительный интервал для математического ожидания твердости на поверхности катания головки рельса составил $\delta = 24,45$ ($\varepsilon = 9,85/24,45 \approx 0,4$), и если бы нам было необходимо определить твердость с точностью $\pm 10HV$ ($\varepsilon \approx 1$), то для этого потребовалось бы еще, как минимум, четыре измерения (кроме уже трех имеющихся). Действительно, при $\varepsilon = 1$ и $P=0,95$ ($\alpha=0,05$), как уже было отмечено, по (3.38) получается $n \geq z_{1-0,05/2}^2 \cdot (1)^2 = (1,96)^2 \approx 4$, затем при $m = 4-1 = 3$ $t_{0,05,2} \approx 3$ (**СТЬЮДРАСПОБР**(0,05;3) = 3,182449), по (3.27) получаем $n \geq t_{0,05,3}^2 \cdot 1^2 \approx 9$; на следующей итерации $t_{0,05,8} \approx 2,3$ (**СТЬЮДРАСПОБР**(0,05;8) = 2,306006), $n \geq 2,3^2 \approx 5$ и затем $t_{0,05,4} \approx 2,8$ (**СТЬЮДРАСПОБР**(0,05;4) = 2,776451), $n \geq 2,8^2 \approx 7$.

Количество опытов, необходимых для построения доверительных интервалов для математического ожидания при некоторых других δ/S_x и P , приведены в табл. 3.1 (для $P=0,95$ в скобках приведены значения, рассчитанные по формуле (3.37)).

Таблица 3.1

Необходимое количество измерений при построении доверительного интервала для математического ожидания

δ/S_x	$P=0,90$	$P=0,95$	$P=0,99$
1	5	7 (4)	11
0,5	13	19 (16)	31
0,4	19	27 (24)	46
0,3	32	46 (48)	78
0,1	273	387 (384)	668

3.3. Статистические гипотезы

Как уже видно из изложенного выше материала, при статистическом оценивании (т.е. при приближенном определении случайной величины) для обос-

нования (состоятельности, несмещенности и эффективности) выбора той или иной оценки неизвестного параметра распределения приходится высказывать предположение, что, например, случайная величина не противоречит нормальному закону распределения. Кроме того, использование всех имеющихся выборочных значений при расчете оценок, так или иначе, предполагает, что среди них нет грубых ошибок (резко выделяющихся значений). С еще большим количеством различных предположений (гипотез) приходится сталкиваться, когда необходимо не только определять случайные величины, но и сравнивать их между собой, и тем более, когда по результатам эксперимента строится функция отклика.

Статистическая гипотеза – любое предположение, касающееся неизвестного распределения случайной величины.

Например, специалиста интересует, удалось ли добиться повышения механической прочности окатышей при использовании новой технологии их обжига. Он может сформулировать следующую гипотезу: "Механическая прочность окатышей увеличилась". Эта гипотеза (нулевая гипотеза) будет подлежать проверке в ходе проведения опытов. Кроме того, можно сформулировать и любую другую (альтернативную) гипотезу, например: "Изменения механической прочности не произошло" или "Механическая прочность окатышей, наоборот, даже уменьшилась".

Процесс принятия решения называется проверкой статистической гипотезы. Поскольку мы выдвигали гипотезу, опираясь только на случайные выборочные значения, наши выводы будут носить вероятностный характер, то есть мы не можем дать точного ответа: да или нет. Можно будет лишь с некоторой долей уверенности (с некоторой вероятностью) утверждать, что данные не противоречат (или противоречат) предположению.

Статистические гипотезы можно разделить на следующие группы.

1. Гипотезы о параметрах распределения. Эти гипотезы представляют собой предположение о значении некоторых параметров распределения генеральной совокупности. Пусть, например, высказывается гипотеза о том, что параметры (математическое ожидание, дисперсии) в двух выборках равны между собой. Обычно гипотезы о параметрах

распределения можно выдвигать, располагая достаточно большой информацией о генеральной совокупности или имея весомые основания считать известным ее закон распределения.

2. Гипотезы о виде распределения. Это более общие гипотезы, они выдвигаются в условиях недостаточной информации о генеральной совокупности. По выборке выдвигается гипотеза о том, соответствуют ли данные, например, нормальному закону распределения. Заметим, что проверка гипотезы о нормальности распределения может помочь при дальнейшей обработке выборки: если случайную величину достаточно уверенно можно считать нормально распределенной, то к ней применимы все теоремы о нормальных величинах, в частности имеется возможность построить доверительные интервалы для параметров.

Нулевая гипотеза H_0 – гипотеза, подлежащая проверке. Это гипотеза, имеющая наиболее важное значение в проводимом исследовании. Нулевую гипотезу выдвигают и затем проверяют с помощью статистических критериев с целью выявления оснований для ее отклонения и принятия альтернативной гипотезы.

Альтернативная гипотеза H_1 – каждая допустимая гипотеза, отличная от нулевой. Обычно в качестве альтернативной гипотезы принимают гипотезу вторую по значимости после основной.

Предположение, которое касается неизвестного параметра распределения, когда вид распределения известен (например, закон Гаусса), называется параметрической гипотезой, а предположение, при котором вид распределения неизвестен, называется непараметрической гипотезой.

Задача исследователя заключается в том, чтобы на основе анализа опытных данных, полученных по выборке, принять ту или иную гипотезу относительно свойств генеральной совокупности, используя при этом какой-либо способ (критерий) проверки высказанного предположения.

Статистический критерий – однозначно определенный способ проверки статистических гипотез.

Критерии для проверки параметрических гипотез называются параметрическими, а для проверки непараметрических гипотез - соответственно непараметрическими.

Естественно, что прежде чем использовать тот или иной параметрический критерий, экспериментатор должен найти способ убедиться в том, согласуется или нет распределение исследуемой им случайной величины с тем или иным теоретическим (например, нормальным) распределением.

Критерий согласия – статистический критерий для проверки гипотезы о согласии (равенстве) распределения случайной величины исследуемой совокупности с теоретическим распределением или гипотезы о согласии распределений в двух и более совокупностях.

Как и при статистическом оценивании, любой критерий может быть построен только на основе тех результатов наблюдений, которые имеются в распоряжении исследователя, т.е. путем вычисления той или иной статистики. А как уже ранее было отмечено, любая статистика как некоторая функция случайной величины (функция от результатов наблюдений) также является случайной величиной.

Таким образом, статистические гипотезы всегда носят вероятностный характер. Это говорит о том, что, основываясь на той или иной статистике и принимая нулевую гипотезу в качестве рабочей (либо отвергая эту гипотезу и в качестве рабочей принимая альтернативную), исследователь может совершить ошибки. Ситуации, возникающие при проверке статистических гипотез, представлены в табл.3.2.

1. *Гипотеза H_0 верна, и она не отвергается*, т.е. принятое решение отражает истинное положение и принимается верная гипотеза.
2. *Гипотеза H_0 верна, но она отвергается*, т.е. в этом случае допущена ошибка первого рода. Ошибка первого рода – ошибка, заключающаяся в том, что отвергают нулевую гипотезу, в то время как в действительности эта гипотеза верна. Вероятность этого события по определению равна уровню значимости α . Уровень значимости α – вероятность ошибки первого рода. Так как уровень значимости задается произвольно, можно

снизить вероятность ошибки первого рода до сколь угодно низкого уровня.

3. *Гипотеза H_0 неверна, и она отвергается.* Опять принятое решение отражает истинное положение и отвергается неверная гипотеза.
4. *Гипотеза H_0 неверна, но она не отвергается.* В этом случае допущена ошибка второго рода. Ошибка второго рода – ошибка, заключающаяся в том, что принимают нулевую гипотезу, в то время как в действительности эта гипотеза неверна. Если вероятность ошибки второго рода обозначить как β , то величина $1 - \beta$ носит название мощность критерия. Мощность критерия – вероятность того, что если верна альтернативная гипотеза, то нулевая гипотеза будет отвергнута.

Таблица 3.2

Возможные исходы при проверке статистических гипотез

Фактическая ситуация	H_0 - принимается	H_0 - отвергается
H_0 - верна	Правильное решение	Ошибка первого рода (α)
H_0 – не верна	Ошибка второго рода (β)	Правильное решение

Значения применяемой для данного критерия статистики, при которых для выбранного уровня значимости отвергается нулевая гипотеза, образуют так называемую критическую область.

Критическая область ω – область со следующими свойствами: если значения применяемой статистики принадлежат данной области, то отвергают нулевую гипотезу; в противном случае ее принимают.

Приведенные определения намечают самую простую форму проверки статистических гипотез. Для того чтобы пояснить сущность этого метода, предположим, что выборочная величина Θ , представляющая собой несмещенную оценку параметра Θ_0 , имеет плотность распределения $f(\Theta)$. Если гипотеза, состоящая в том, что $\Theta = \Theta_0$, верна, то функция $f(\Theta)$ должна попадать в среднюю область, как показано на рис.3.4. Вероятность того, что параметр Θ не будет превышать нижнего уровня $\Theta_{1-\alpha/2}$, составит

$$P(\Theta \leq \Theta_{1-\alpha/2}) = \int_{-\infty}^{\Theta_{1-\alpha/2}} f(\Theta) d\Theta = \alpha/2.$$

Вероятность того, что параметр Θ превысит верхний уровень $\Theta_{\alpha/2}$, равна

$$P(\Theta > \Theta_{\alpha/2}) = \int_{\Theta_{\alpha/2}}^{\infty} f(\Theta) d\Theta = \alpha/2.$$

Следовательно, вероятность того, что параметр Θ выйдет за пределы интервала $[\Theta_{1-\alpha/2}; \Theta_{\alpha/2}]$, составляет α . Теперь примем величину α малой, чтобы попадание параметра Θ за пределы интервала $[\Theta_{1-\alpha/2}; \Theta_{\alpha/2}]$ было маловероятно. Если после извлечения выборки и определения величины Θ окажется, что она выходит за пределы интервала $[\Theta_{1-\alpha/2}; \Theta_{\alpha/2}]$ и попадает в критическую область, то в этом случае есть серьезные основания подвергнуть сомнению справедливость проверяемой гипотезы $\Theta = \Theta_0$. С другой стороны, если параметр Θ попадает в интервал $[\Theta_{1-\alpha/2}; \Theta_{\alpha/2}]$, то в этом случае нет серьезных оснований подвергать сомнению справедливость проверяемой гипотезы, и гипотезу равенства $\Theta = \Theta_0$ можно принять. Как видно из рис.3.4, ошибка первого рода допускается, если гипотеза верна, а параметр Θ попадает в область отклонения гипотезы. Отсюда следует, что вероятность допустить ошибку первого рода равна α , т.е. уровню значимости критерия.

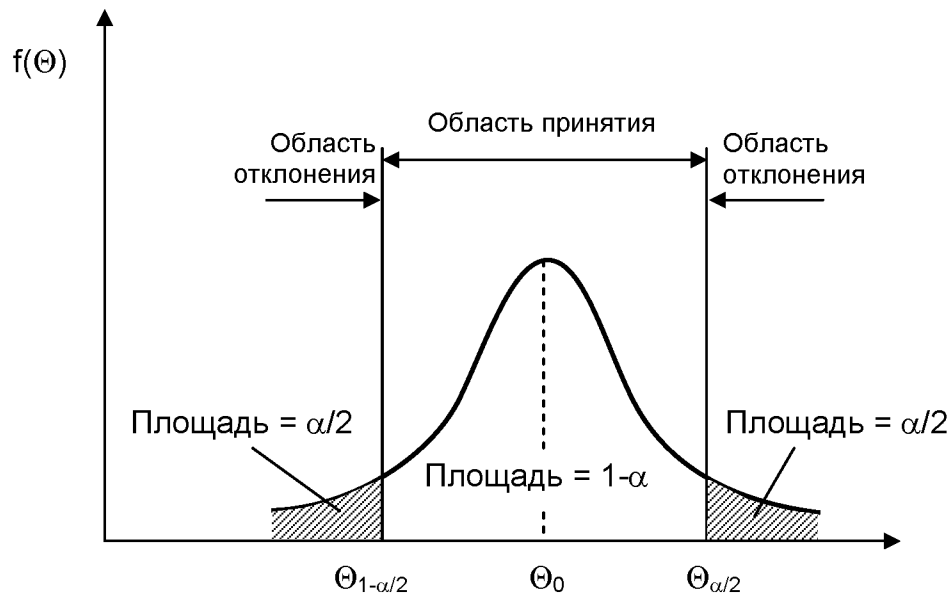


Рис. 3.4. Области принятия и отклонения гипотезы при проверке гипотез

Для того чтобы найти, какова вероятность допустить ошибку второго рода, необходимо задать определенную величину отклонения истинного значения от гипотетического значения параметра, которое требуется определить. Предположим, например, что истинное значение параметра в действительности равно Θ_0+d или Θ_0-d , как показано на рис.3.5. Если, согласно гипотезе, $\Theta=\Theta_0$, а в действительности $\Theta=\Theta_0 \pm d$, то вероятность того, что Θ попадает в область принятия гипотезы, т.е. в интервал $[\Theta_{1-\alpha/2}; \Theta_{\alpha/2}]$, составляет β . Это означает, что вероятность допустить ошибку второго рода при выявлении отклонения $\pm d$ от гипотетического значения Θ равна β .

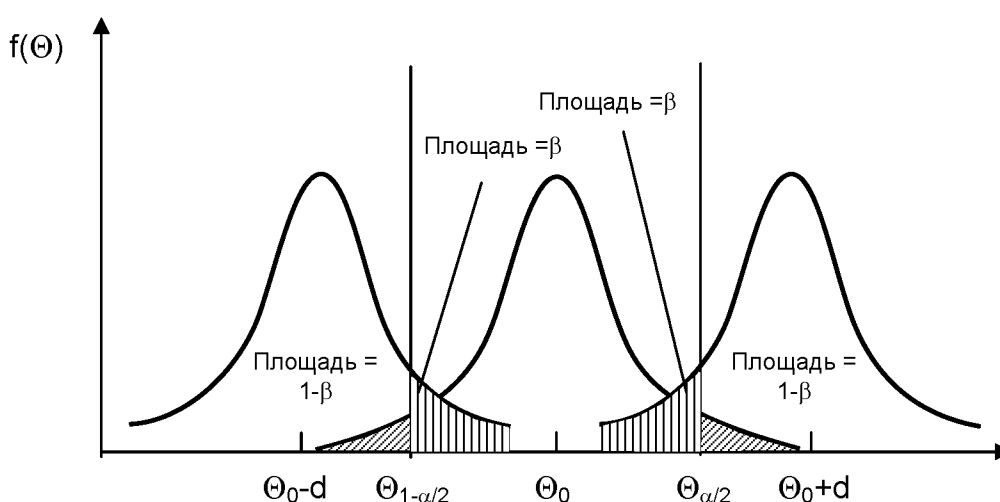


Рис. 3.5. Области принятия и отклонения гипотезы, соответствующие ошибке второго рода при проверке гипотезы

Очевидно, что при любом заданном объеме выборки вероятность допустить ошибку первого рода можно сократить, уменьшив уровень значимости α . Однако при этом увеличивается вероятность допущения ошибки второго рода (снижается мощность критерия). Таким образом, в большинстве случаев нельзя добиться минимального значения вероятностей α и β одновременно. Поступают обычно следующим образом: фиксируют вероятность α ошибки первого рода, а затем добиваются минимума вероятности β ошибки второго рода. За счет чего можно уменьшить β при фиксированном значении α ? За счет правильного выбора критической области: при заданной альтернативе H_1 критическую область выбирают таким образом, чтобы значение β (вероятность принять неверную гипотезу) было наименьшим из возможных. Таким образом, за-

дача состоит в построении наиболее мощного критерия $(1 - \beta)$ при заданном уровне значимости α .

Различают односторонние и двусторонние критические области. Различные варианты областей представлены на рис. 3.6.

Если хотят убедиться, что одна случайная величина строго больше или строго меньше другой, то используют одностороннюю критическую область (рис.3.6.а,б).

Таким образом, для данного случая $H_0: \Theta = \Theta_0$;

$H_1^{(1)} : \Theta < \Theta_0$;

$H_1^{(2)} : \Theta > \Theta_0$.

Если проверяют как положительные, так и отрицательные расхождения между изучаемыми величинами, то используют двусторонние критические области (рис.3.6,в)

$H_0 : \Theta = \Theta_0$;

$H_1^{(3)}: \Theta \neq \Theta_0$.

Подводя итог всему вышесказанному, алгоритм проверки любой статистической гипотезы в самом общем случае заключается в следующем:

- 1) формулирование нулевой гипотезы H_0 ;
- 2) выбор одной из альтернативных гипотез $H_1^{(1)}$, $H_1^{(2)}$, $H_1^{(3)}$;
- 3) поиск критерия, по которому может быть проверена сформулированная нулевая гипотеза H_0 ;
- 4) расчет значения статистики, применяемой для данного критерия;
- 5) выбор уровня значимости α ;
- 6) построение критической области ω при выбранном уровне значимости α ;
- 7) принятие решения: если значение статистики попало в критическую область – нулевая гипотеза отвергается, при этом вероятность ошибки (первого рода) не превышает выбранный уровень значимости; в противном случае – нулевая гипотеза принимается.

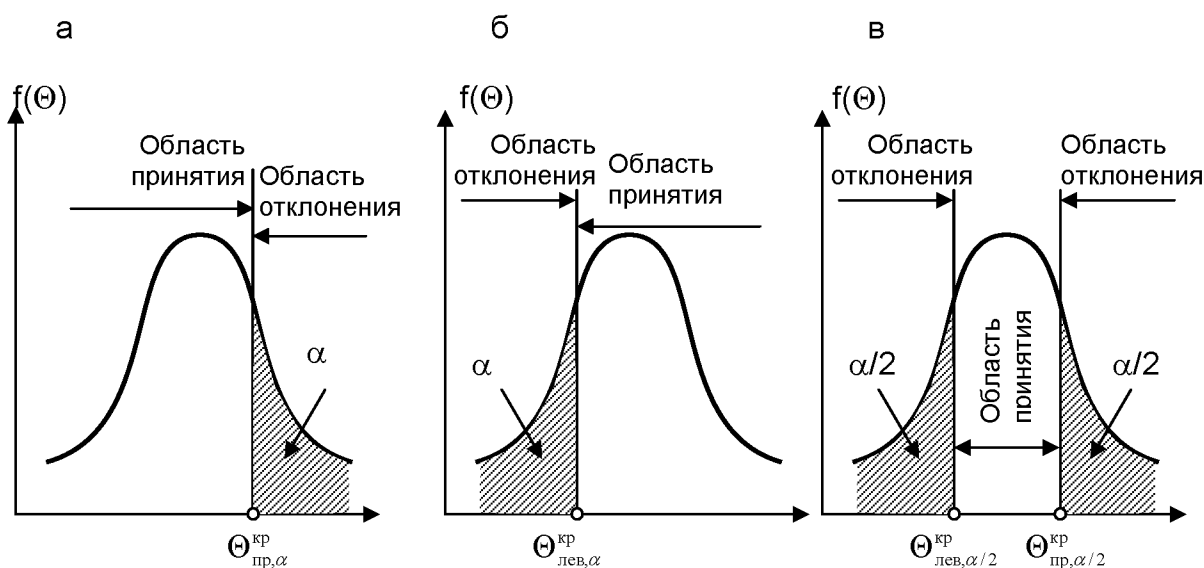


Рис. 3.6. Критические области плотности распределения:
 а – правосторонняя, б – левосторонняя, в – двусторонняя

При использовании механизма статистических гипотез следует помнить, что даже в случае принятия нулевой гипотезы в $100\alpha\%$ вывод будет ошибочным в связи со всегда имеющейся вероятностью совершить ошибку первого рода. Причем если значение статистики не попадает в критическую область, то прежде, чем принять нулевую гипотезу, необходимо оценить вероятность ошибки второго рода, т.е. рассчитать мощности критерия. Если же его величина оказывается недостаточной для решения поставленной задачи, требуется увеличение объема опытных данных (однако поскольку при обработке эксперимента исследователи зачастую уже не имеют возможности увеличить объем выборки, то они обычно пропускают данный пункт).

3.4. Отсев грубых погрешностей

Часто даже тщательно поставленные эксперименты могут давать неоднородные данные, поскольку в процессе эксперимента могут измениться условия проведения опытов. Если экспериментатор по каким-либо причинам не уловил этих изменений, наблюдения, соответствующие разным уровням факторов, будут принадлежать к разным генеральным совокупностям. Данные, соответствующие изменившимся условиям, называют грубыми погрешностями (ошибками) или резко выделяющимися (аномальными) значениями. Грубые погрешности появляются также при неправильной записи показаний приборов.

В литературе приводятся сведения о том, что экспериментальные данные могут содержать ~ 10% аномальных значений. Однако эти 10% могут дать сильное смещение при оценке параметров распределения, особенно для дисперсии, так как ошибки заметно отклоняются от основной группы значений, а на дисперсию особенно сильно влияют крайние члены вариационного ряда (вариационный ряд – результаты наблюдений, расположенные в возрастающей последовательности $x_1 \leq x_2 \leq x_3 \dots \leq x_i \dots \leq x_n$).

В случае отсева грубых погрешностей (ошибок) нулевая гипотеза формулируется следующим образом:

H_0 : "Среди результатов наблюдений (выборочных, опытных данных) нет резко выделяющихся (аномальных) значений".

Альтернативной гипотезой может быть

либо $H_1^{(1)}$: "Среди результатов наблюдений есть только одна грубая ошибка",

либо $H_1^{(2)}$: "Среди результатов наблюдений есть две или более грубых ошибки".

В литературе можно встретить большое количество различных критериев для отсева грубых погрешностей наблюдений. Обычно экспериментаторы имеют дело с выборками небольшого объема (т.е. когда генеральная дисперсия σ_x^2 неизвестна и оценивается по опытным данным через выборочную дисперсию S_x^2), причем именно в этом случае аномальные данные имеют большой вес. Наиболее распространенными и теоретически обоснованными в этом случае являются критерий Н.В. Смирнова (используется при $H_1^{(1)}$) и критерий Диксона (применим как при $H_1^{(1)}$, так и при $H_1^{(2)}$).

3.4.1. Критерий Н.В. Смирнова

Если известно, что есть только одно аномальное значение (альтернативная гипотеза $H_1^{(1)}$), то оно будет крайним членом вариационного ряда. Поэтому проверять выборку на наличие одной грубой ошибки естественно при помощи статистики

$$u_1 = \frac{\bar{x} - x_1}{S_x}, \quad (3.39)$$

если сомнение вызывает первый член вариационного ряда $x_1 = \min_i x_i$,

или

$$u_n = \frac{x_n - \bar{x}}{s_x}, \quad (3.40)$$

если сомнителен максимальный член вариационного ряда $x_n = \max_i x_i$.

Этот критерий впервые был предложен Н.В. Смирновым. Он исследовал распределение статистик [(3.39), (3.40)] и составил таблицы процентных точек $u_{\alpha,n}$ (квантили порядка $p = 1 - \alpha$) для $\alpha = 0,1; 0,05; 0,01$ при $3 \leq n \leq 20$ [11].

При выбранном уровне значимости α критическая область для критерия Н.В. Смирнова строится следующим образом:

$$u_1 > u_{\alpha,n} \text{ или } u_n > u_{\alpha,n}, \quad (3.41)$$

где $u_{\alpha,n}$ – это табличные значения (см. [6] или табл. П.7).

В случае если выполняется последнее условие (статистика попадает в критическую область), то нулевая гипотеза отклоняется, т.е. выброс x_1 или x_n не случаен и не характерен для рассматриваемой совокупности данных, а определяется изменившимися условиями или грубыми ошибками при проведении опытов. В этом случае значение x_1 или x_n исключают из рассмотрения, а найденные ранее оценки подвергаются корректировке с учетом отброшенного результата.

3.4.2. Критерий Диксона

В критерии Диксона применяется статистика:

- если подозрительная «чужеродная» точка имеет наибольшее значение,

$$r_{i,j} = \frac{x_n - x_{n-i}}{x_n - x_{j+1}}, \quad (3.42)$$

- если подозрительная «чужеродная» точка имеет наименьшее значение,

$$r_{i,j} = \frac{x_{1+i} - x_1}{x_{n-j} - x_1}, \quad (3.43)$$

где x_n, x_{n-i}, x_{j+1} – члены вариационного ряда $x_1 \leq x_2 \leq x_3 \dots \leq x_i \dots \leq x_n$.

Диксоном были получены распределения для $r_{10}, r_{11}, r_{12}, r_{20}, r_{21}$ и r_{22} и построены таблицы для $\alpha = 0,1; 0,05; 0,01$ и $0,005$ при $3 \leq n \leq 30$ [11].

Статистика

$$r_{10} = \frac{x_n - x_{n-1}}{x_n - x_1}$$

используется для проверки максимального или минимального члена вариационного ряда (одна грубая ошибка, альтернативная гипотеза $H_1^{(1)}$) при $3 \leq n \leq 7$. Если при том же объеме выборки предполагается наличие двух и более резко выделяющихся значений (альтернативная гипотеза $H_1^{(2)}$), то используется статистика r_{20} .

Статистики критерия Диксона, используемые при других объемах выборки, приведены в табл. 3.3.

Таблица 3.3

Статистики критерия Диксона, используемые при различных объемах выборки n

Объем выборки n	Число грубых погрешностей	
	одна	две и более
3...7	r_{10}	r_{20}
8...10	r_{11}	r_{20}
11...13	r_{21}	r_{21}
14...30	r_{22}	r_{22}

Критическая область в критерии Диксона выглядит аналогично критерию Н.В. Смирнова и включает значения

$$r_{ij} > (r_{ij})_{\alpha,n}, \quad (3.44)$$

где $(r_{ij})_{\alpha,n}$ – табличные значения (см. [11] или табл. П.8).

Рассмотрим небольшой пример.

Пример 3.3. Пирометром измеряется температура поверхности нагретого тела (например, прокатываемой заготовки, причем будем предполагать, что

температура ее видимой поверхности во всех точках одинакова). Было проведено шесть измерений температуры T °С, и получены следующие значения: 925, 930, 950, 975, 990, 1080 ($n = 6$, причем, как видно, все значения приведены в возрастающей последовательности, т.е. в виде вариационного ряда $T_1=925 \leq T_2=930 \leq T_3=950 \dots \leq T_6=1080$). Можно ли значение $T_6=1080$ считать грубой погрешностью, полученной, допустим, в результате неправильной регистрации показаний пирометра?

Для ответа на поставленный в этом примере вопрос предварительно вычислим оценки параметров распределения исследуемой случайной величины T (предполагая, что она не противоречит нормальному закону распределения): выборочное среднее арифметическое \bar{T} и выборочное среднее квадратичное отклонение S_T :

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = (925 + 930 + 950 + 975 + 990 + 1080) / 6 = 975;$$

$$S_T^2 = \frac{1}{n-1} \left[\sum_{i=1}^n T_i^2 - \frac{1}{n} \left(\sum_{i=1}^n T_i \right)^2 \right] = \frac{1}{6-1} [(925^2 + 930^2 + 950^2 + 975^2 + 990^2 + 1080^2) - \frac{1}{6} (925 + 930 + 950 + 975 + 990 + 1080)^2] = 3280;$$

$$S_T = +\sqrt{S_T^2} = +\sqrt{3280} = 57,27.$$

В электронных таблицах Microsoft Excel для этих расчетов можно было бы использовать две статистические функции **СРЗНАЧ** (925;930;950;975;990;1080) =975 и **СТАНДОТКЛОН** (925;930;950;975;990;1080)= 57,27128.

Теперь воспользуемся предложенным выше алгоритмом проверки статистических гипотез.

1. Формулируем нулевую гипотезу H_0 : "Среди значений 925; 930; 950; 975; 990; 1080 нет грубых погрешностей".
2. Исходя из условий примера 3.2, выбираем следующую альтернативную гипотезу $H_1^{(1)}$: "Значение 1080 является (одной) грубой погрешностью".
3. Сформулированная нулевая гипотеза H_0 может быть проверена по любому из приведенных в этом разделе критериев, т.е. как по критерию Н.В. Смир-

нова, так и по критерию Диксона (хотя в литературе могут быть найдены и другие критерии). Для начала остановимся на критерии Н.В. Смирнова.

4. Значение статистики критерия Н.В. Смирнова в примере 3.2 равно (см. (3.40))

$$u_6 = \frac{T_6 - \bar{T}}{s_T} = \frac{1080 - 975}{57,27} = 1,83.$$

5. Уровень значимости α примем равным 0,05.
6. По табл. П.7 при $\alpha = 0,05$ и $n = 6$ находим $u_{0,05;6} = 1,82$, и с использованием (3.41) строим критическую область $\omega: u_6 > u_{0,05;6}$, т.е. $u_6 > 1,82$.
7. Принимаем решение: поскольку значение статистики ($1,83 > 1,82$) попало в критическую область – нулевая гипотеза отвергается, и в качестве рабочей принимается альтернативная гипотеза, т.е. значение 1080 с вероятностью 0,95 (уровень значимости, не превышает 0,05) по критерию Н.В. Смирнова можно считать грубой погрешностью.

Интересно отметить, что если бы на этапе 5 мы приняли $\alpha = 0,01$, по таблицам критерия Н.В. Смирнова $u_{0,01;6} = 1,94$ и подсчитанное значение статистики при этом уровне значимости, то оно не попало бы в критическую область ($1,83 < 1,94$). Следовательно, при $\alpha = 0,01$ мы не можем отвергнуть нулевую гипотезу, т.е. по критерию Н.В. Смирнова с вероятностью 0,99 (надежностью, достоверностью) мы не можем сказать, что значение 1080 является грубой погрешностью.

В завершение данного примера рассмотрим, как бы выглядели наши расчеты, если на этапе 3 мы бы остановились на критерии Диксона.

4. При $n = 6$ и альтернативной гипотезе, что имеется только одна грубая погрешность, в критерии Диксона используется статистика r_{10} (см. табл. 3.3), значение которой в примере 3.2 (см. (3.43)):

$$r_{10} = \frac{T_6 - T_{6-1}}{T_6 - T_{0+1}} = \frac{T_6 - T_5}{T_6 - T_1} = \frac{1080 - 990}{1080 - 925} = 0,581.$$

5. Уровень значимости α примем равным 0,05.
6. По табл. П.8 при $\alpha = 0,05$ и $n = 6$ находим $(r_{10})_{0,05;6} = 0,560$, и с использованием (3.44) строим критическую область $\omega: r_{10} > (r_{10})_{0,05;6}$, т.е. $r_{10} > 0,560$.

7. Принимаем решение: поскольку значение статистики ($0,581 > 0,560$) попало в критическую область – нулевая гипотеза отвергается, и в качестве рабочей принимается альтернативная, т.е. значение 1080 с вероятностью 0,95 и по критерию Диксона можно считать грубой погрешностью.

Заметим, однако, как и по критерию Н.В. Смирнова, высказать подобное утверждение с вероятностью 0,99 по критерию Диксона мы не имеем права, поскольку по таблицам (r_{10}) $0,01;6 = 0,698$.

3.5. Сравнение двух рядов наблюдений

При проведении и анализе результатов экспериментальных исследований часто приходится сравнивать две партии изделий, показания двух или нескольких приборов, анализировать результаты работы однотипных агрегатов, сравнивать результаты исследований двух проб материалов и т.д. Вот некоторые примеры подобных ситуаций.

1. Необходимо сравнить показания двух приборов, измеряющих одну и ту же величину, когда этими рабочими средствами измерений получено два ряда наблюдений данной величины. Одинакова ли точность измерения одного и того же технологического параметра разными приборами?

2. Требуется поверить рабочее средство измерения (т.е. определить, не выходят ли погрешности его измерений за пределы регламентированных значений) с помощью образцового средства измерения. Равно ли математическое ожидание показаний данного прибора действительному значению измеряемого параметра?

3. Два агрегата выпускают одну и ту же продукцию. Необходимо сделать вывод о том, какой из них лучше или хуже в каком-либо смысле.

Решение подобных задач осуществляется также с использованием аппарата проверки статистических гипотез. Ведь если нам необходимо было бы сравнить две случайные величины X и Y , имеющие нормальное распределение, при известных их математических ожиданиях и дисперсиях M_x ; σ_x^2 и M_y ; σ_y^2 , то вопрос, очевидно, решался бы достаточно просто. Две случайные вели-

чины с нормальным распределением равны между собой (имеют одинаковое распределение, т.е. имеют одну и ту же функцию распределения $F(X) = F(Y)$ или плотность распределения $f(X) = f(Y)$), когда равны между собой их математические ожидания ($M_x = M_y$) и дисперсии ($\sigma_x^2 = \sigma_y^2$), поскольку только эти два параметра полностью определяют нормальное (двухпараметрическое) распределение (см. (2.12) или (2.21)).

Однако, как это уже неоднократно ранее отмечалось, любой из параметров распределения случайной величины Θ может быть найден лишь по всей генеральной совокупности, т.е. только теоретически при проведении бесконечно большого количества опытов. Практически, по выборке ограниченного объема, исследователь может определить только приближенное значение параметра – его оценку Θ^* . При этом вероятность того, что оценка Θ^* совпадет со значением оцениваемого параметра Θ , очень мала. Следовательно, даже если равны между собой параметры распределений двух случайных величин ($\Theta_x = \Theta_y$), то их оценки скорее всего не будут одинаковыми ($\Theta_x^* \neq \Theta_y^*$).

Поэтому при сравнении двух случайных величин обычно приходится высказывать и проверять нулевую гипотезу $H_0: \Theta_x = \Theta_y$, при альтернативных гипотезах типа $H_1^{(1)}: \Theta_x < \Theta_y$ или $H_1^{(2)}: \Theta_x > \Theta_y$, $H_1^{(3)}: \Theta_x \neq \Theta_y$,

3.5.1. Сравнение двух дисперсий

При выполнении измерений в различных условиях часто возникает задача сравнения степени разброса (дисперсий) исследуемых параметров (случайных величин).

Проверка гипотезы о равенстве дисперсий имеет большое значение, так как измеряемая дисперсией величина рассеяния характеризует такие исключительно важные показатели, как точность машин, приборов, стабильность технологических процессов, качество готовой продукции и т.д. Поэтому, например, о преимуществах той или иной технологии или о качестве выпускаемой продукции вывод можно часто сделать в результате сравнения дисперсий тех параметров, которые их характеризуют.

Таким образом, требуется установить, являются ли выборочные дисперсии $S_1^2 \neq S_2^2$ со степенями свободы m_1 и m_2 значимо отличающимися или же они характеризуют выборки, взятые из одной и той же генеральной совокупности или из генеральных совокупностей с равными дисперсиями ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). В этом случае нулевая гипотеза формулируется в виде $H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2$, т.е. между двумя генеральными дисперсиями различия нет при заданном уровне значимости α .

Для проверки этой гипотезы используется критерий, основанный на распределении Фишера, зависящем только от числа степеней свободы m_1 и m_2 . Аналитическое выражение критерия Фишера имеет вид

$$F = (S_1^2 / \sigma_1^2) / (S_2^2 / \sigma_2^2) = (S_1^2 / S_2^2) / (\sigma_2^2 / \sigma_1^2). \quad (3.44a)$$

Плотность распределения величины F_{m_1, m_2} , представленная на рис. 3.7, есть функция

$$f(F) = \begin{cases} \frac{\Gamma\left(\frac{m_1 + m_2}{2}\right) \left(\frac{m_1}{m_2}\right)^{m_1/2} F^{\left(\frac{m_1}{2} - 1\right)} \Gamma\left(\frac{m_1}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right) \Gamma\left(\frac{m_2}{2}\right) \left[1 + \frac{m_1 F}{m_2}\right]^{(m_1 + m_2)/2}} & \text{при } F \geq 0; \\ 0 & \text{при } F < 0. \end{cases} \quad (3.44б)$$

Надо иметь в виду, что скорость возрастания и убывания функции, а также величина и положение максимума зависят от параметров m_1 и m_2 .

Соответствующая функция распределения величины F_{m_1, m_2} определяется через плотность распределения

$$F(F) = \int_{-\infty}^F f(\xi) d\xi. \quad (3.44в)$$

Существуют статистические таблицы как с табулированными значениями функции распределения Фишера для принятого уровня значимости, так и с табулированными значениями квантилей этого распределения (см. табл. П.4 и П.5).

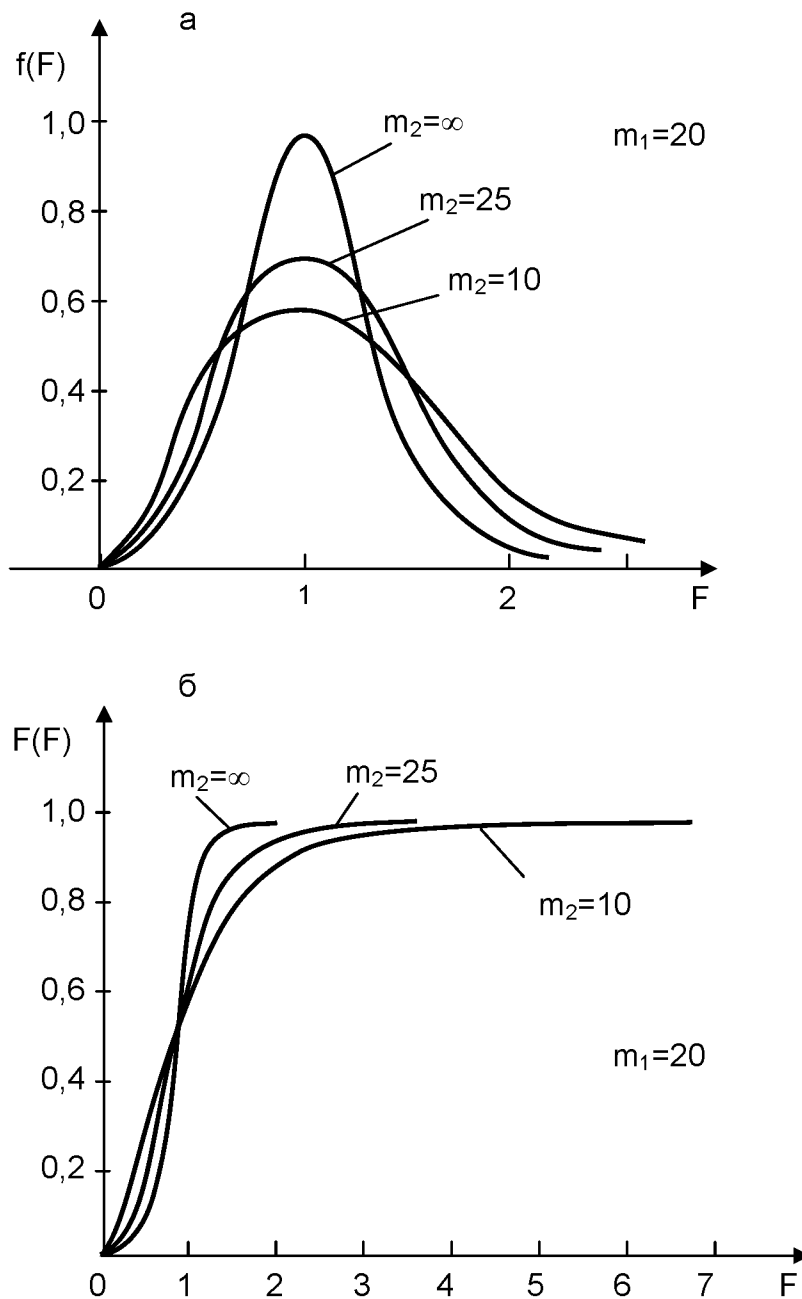


Рис. 3.7. Плотность (а) и функция (б) F -распределения (частный случай при $m_1=20$)

Поскольку по условию нуль-гипотезы $\sigma_1^2 = \sigma_2^2$, то выражение можно представить как отношение выборочных дисперсий

$$F = S_1^2 / S_2^2,$$

где $S_1^2 > S_2^2$.

Если при проверке нулевой гипотезы $H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2$ альтернативной является гипотеза $H_1^{(1)}: \sigma_1^2 > \sigma_2^2$, то применяют одностороннее неравенство

$$F = S_1^2 / S_2^2 > F_{\alpha, m_1, m_2}.$$

Для альтернативной гипотезы $H_1^{(2)}: \sigma_1^2 \neq \sigma_2^2$, когда соотношение между генеральными дисперсиями неизвестно, различие между дисперсиями считают значимым, если выполняется условие

$$F = S_1^2 / S_2^2 > F_{(\alpha/2), m_1, m_2}.$$

Таким образом, алгоритм решения задачи сводится к следующему.

Пусть по результатам испытаний двух независимых выборок объемом n_1 и n_2 из нормально распределенных совокупностей подсчитаны оценки дисперсий S_1^2 и S_2^2 , причем $S_1^2 > S_2^2$. Требуется проверить предположение (нулевую гипотезу H_0) о том, что указанные выборки принадлежат генеральным совокупностям с равными дисперсиями.

В соответствии с общим алгоритмом проверки любой статистической гипотезы:

1. $H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2$.
2. Возможно два варианта альтернативной гипотезы:

$$H_1^{(1)}: \sigma_1^2 \neq \sigma_2^2;$$

$$H_1^{(2)}: \sigma_1^2 > \sigma_2^2.$$

Предположить вариант альтернативной гипотезы $H_1^{(3)}: \sigma_1^2 < \sigma_2^2$, конечно же, возможно, но вряд ли целесообразно при условии, что $S_1^2 > S_2^2$.

3. Используется F-критерий (критерий Фишера) – это отношение двух дисперсий (большой к меньшей), F - статистика поэтому имеет вид

$$F = \frac{S_1^2}{S_2^2}, \tag{3.45}$$

где $S_1^2 > S_2^2$.

Очевидно, что значения F всегда больше единицы.

4. Выбирается уровень значимости α .

5. Границы критической области можно установить по таблицам квантилей F - распределения (см. [11] или табл. П.4, П.5, а в Microsoft Excel для этого используется функция **ФРАСПОБР**) для числа степеней свободы $m_1 = n_1 - 1$ и $m_2 = n_2 - 1$ и уровня значимости

- при альтернативной гипотезе $H_1^{(1)}$: $\sigma_1^2 \neq \sigma_2^2$ уровень значимости равен $\alpha/2$ и критическая область определяется соотношением $F > F_{(\alpha/2), m_1, m_2}$;
- при альтернативной гипотезе $H_1^{(2)}$: $\sigma_1^2 > \sigma_2^2$ уровень значимости равен α и критическая область определяется соотношением $F > F_{\alpha, m_1, m_2}$.

6. Нулевую гипотезу принимают, т.е. полагают, что $\sigma_1^2 = \sigma_2^2 = \sigma^2$ при выполнении одного из неравенств (для различных альтернативных гипотез):

- $F \leq F_{(\alpha/2), m_1, m_2}$ при $H_1^{(1)}$ $\sigma_1^2 \neq \sigma_2^2$;
- $F \leq F_{\alpha, m_1, m_2}$ при $H_1^{(2)}$ $\sigma_1^2 > \sigma_2^2$.

В случае подтверждения нулевой гипотезы, по двум выборочным дисперсиям производят оценку общей генеральной дисперсии σ^2

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \quad (3.46)$$

которая может быть использована для дальнейшего анализа опытных данных.

Проиллюстрируем применение критерия Фишера на следующем примере.

Пример 3.4. Проводятся измерения одной и той же физической величины (температуры, давления, состава газа и т.п.). Первым (старым) измерительным прибором выполнено 200 измерений, которые дали выборочную дисперсию $S_1^2 = 3,82$, а вторым (новым) сделано только 15 измерений при выборочной дисперсии $S_2^2 = 2,00$. Можно ли считать, что разброс в показаниях нового прибора существенно ниже, чем у старого?

1. Сформулируем нулевую гипотезу о равенстве дисперсий $H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2$.
2. Выберем альтернативную ей гипотезу $H_1: \sigma_1^2 > \sigma_2^2$.

3. Воспользуемся критерием Фишера и рассчитаем статистику этого критерия $F = 3,82/2,00 = 1,91$.

4. Для уровня значимости $\alpha = 0,05$ строим критическую область при $m_1 = 200-1 = 199$ и $m_2 = 15-1 = 14$; $F_{0,05;199;14} = 2,16$ (см.[11] или **FPAC-ПОБР**(0,05;199;14) = 2,159361).

5. Подсчитанное значение статистики ($F=1,91$) не попадает в критическую область ($1,91 < 2,16$), следовательно, нулевая гипотеза $H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2$ принимается, т.е. по имеющимся экспериментальным данным нет достаточных оснований считать, что результаты измерений нового прибора точнее, чем старого.

Как изменится наш вывод, если мы увеличим число измерений новым прибором до 50 при условии, что выборочная дисперсия его показаний при этом не изменится?

Табличное значение критерия Фишера при этом равно $F_{0,05;199;49} = 1,49$, и значение статистики попадет в критическую область $1,91 > 1,49$, следовательно, в качестве рабочей может быть принята альтернативная гипотеза $H_1: \sigma_1^2 > \sigma_2^2$, т.е. результаты измерений новым прибором точнее, чем старым.

3.5.2. Проверка однородности нескольких дисперсий

Критерий Фишера используется для сравнения только двух дисперсий, однако на практике приходится сравнивать между собой три и более дисперсий.

При сопоставлении дисперсий ряда совокупностей нулевая гипотеза заключается в том, что все k совокупностей, из которых взяты выборки, имеют равные дисперсии.

$$1. H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots = \sigma_k^2 = \sigma^2,$$

т.е. проверке подлежит предположение, что все эмпирические дисперсии $S_1^2, S_2^2, \dots, S_k^2$ относятся к выборкам из совокупности с одной и той же генеральной дисперсией σ^2 .

Пусть среди нескольких серий измерений обнаружена такая, выборочная дисперсия которой S_{\max}^2 заметно больше всех остальных. Задача заключается

в том, чтобы выяснить, можно ли считать отличие выделенной дисперсии S_{\max}^2 существенным. Другими словами, альтернативная гипотеза может быть выбрана как

$$2. H_1: \sigma_{\max}^2 > \sigma^2.$$

3. При равном объеме $n_1 = n_2 = n_3 = \dots = n_k = n$ всех k выборок может быть использован так называемый критерий Кохрена (в ряде книг пишется - Кочрена).

4. Статистика критерия Кохрена G рассчитывается как отношение S_{\max}^2 к сумме всех выборочных дисперсий:

$$G = \frac{S_{\max}^2}{\sum_{i=1}^k S_i^2}. \quad (3.47)$$

5. В дальнейшем для выбранного уровня значимости α определяется табличное значение этого критерия, которое зависит от числа степеней свободы $m = n - 1$ и числа сравниваемых дисперсий $k - G_{\alpha; m; k}$ (см. [11] или табл. П.9).

6. Критическая область строится как $G \geq G_{\alpha; m; k}$.

7. При $G < G_{\alpha; m; k}$ гипотеза $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots = \sigma_k^2 = \sigma^2$ принимается в качестве рабочей, т.е. отличие выделенной дисперсии S_{\max}^2 считается несущественным.

В случае подтверждения однородности дисперсий можно сделать оценку обобщенной дисперсии σ^2 :

$$S^2 = \frac{\sum_{i=1}^k S_i^2}{k}. \quad (3.48)$$

Пример 3.5. Шестью ($k = 6$) приборами произведено по семь измерений ($n = 7$) одного и того же параметра, при этом получены следующие выборочные дисперсии S_i^2 : 3,82; 1,7; 1,3; 0,92; 0,78; 0,81. Можно ли считать, что разброс показаний первого прибора ($S_{\max}^2=3,82$) существенно превышает разбросы показаний остальных пяти приборов?

1. Нулевая гипотеза $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2 = \sigma^2$.

2. Альтернативная гипотеза $H_1: \sigma_{max}^2 > \sigma^2$.

3. Поскольку ($n_1 = n_2 = n_3 = n_4 = n_5 = n_6 = 7$) все шесть выборок имеют одинаковый объем, то может быть использован критерий Кохрена.

4. Значение статистики данного критерия в соответствии с уравнением (3.47) составит:

$$G = \frac{3,82}{3,82 + 1,7 + 1,3 + 0,92 + 0,78 + 0,81} = \frac{3,82}{9,33} = 0,409.$$

5. Табличное значение этого критерия для уровня значимости $\alpha = 0,05$, при числе степеней свободы для каждой из дисперсий $m = 7-1 = 6$ и числе сравниваемых дисперсий $k = 6$, равно $G_{0,05;6;6}=0,418$ (табл.П.9).

6. Так как $G < G_{\alpha;m;n}$, отклонение дисперсии $S_{max}^2 = 3,82$ от остальных нельзя (с вероятностью 0,95) признать существенным, и, следовательно, все дисперсии однородны (т.е. разбросы в показаниях всех шести приборов примерно одинаковы).

Оценка обобщенной дисперсии:

$$S^2 = \frac{\sum_{i=1}^k S_i^2}{k} = \frac{9,33}{6} = 1,56.$$

Критерий Кохрена можно использовать только в тех случаях, когда все сравниваемые дисперсии имеют одинаковое число степеней свободы $m = n - 1$ (одинаковые объемы выборок $n_1 = n_2 = n_3 = \dots = n_k = n$). Если же число измерений n в различных сериях неодинаково, то для проверки однородности дисперсий можно выбрать, например, критерий Бартлетта. При необходимости с процедурой его использования можно познакомиться в литературе по теории вероятности и математической статистике (см. например, [9,10]).

3.5.3. Проверка гипотез о числовых значениях математических ожиданий

Часто для решения вопроса о соответствии произведенной продукции определенным требованиям (например, требованиям ГОСТ или ТУ) при выявлении преимущества того или иного технологического процесса или нового материала и т.д. возникает необходимость по выборочным средним значениям исследуемых случайных величин делать вывод о соответствующих им генеральных значениях математических ожиданий.

При этом может возникнуть задача сравнения неизвестного математического ожидания M_1 , для которого получена оценка через выборочное среднее \bar{x}_1 , с конкретным числовым значением M (например, с известным математическим ожиданием) или задача сравнения двух математических ожиданий M_1 и M_2 , оцененным по двум выборочным средним \bar{x}_1 и \bar{x}_2 .

В первом случае в качестве нулевой гипотезы выдвигается предположение о том, что оцененное математическое ожидание M_1 равно известному математическому ожиданию M .

1. $H_0: M_1 = M$.

2. Альтернативная гипотеза может быть в трех вариантах:

$$H_1^{(1)}: M_1 > M; H_1^{(2)}: M_1 < M; H_1^{(3)}: M_1 \neq M.$$

3. Если генеральная дисперсия σ^2 неизвестна и для нее, по той же самой выборке, что и для \bar{x}_1 , сделана оценка S^2 , то используется t-критерий (распределения Стьюдента).

4. t - статистика имеет вид

$$t = \frac{\bar{x} - M}{S} \cdot \sqrt{n}. \quad (3.49)$$

5. Как и при построении доверительного интервала, для математического ожидания (см. раздел 3.2.1) выбирается уровень значимости α .

6. Для числа степеней свободы $m = n - 1$ (с которым сделана оценка дисперсии) устанавливаются границы критической области по табличным значениям квантилей t-распределения (см., например, [11] или табл. П.6), или их можно определить, воспользовавшись статистической функцией **СТЬЮДРАСПОБР** из электронных таблиц Microsoft Excel.

7. Нулевую гипотезу принимают, т.е. полагают, что $M_1 = M$ при выполнении неравенств:

- для альтернативных гипотез $H_1^{(1)}$: $M_1 > M$ и $H_1^{(2)}$: $M_1 < M$ $|t| \leq t_{2\alpha, m}$;
- для альтернативной гипотезы $H_1^{(3)}$: $M_1 \neq M$ $|t| \leq t_{\alpha, m}$.

Появление в последних неравенствах величин α и 2α при определении табличных значений критерия Стьюдента связано с тем, что обычно эти таблицы (см. табл. П.6) приводятся для двустороннего распределения Стьюдента, т.е. под $t_{\alpha, m}$ понимается величина, которая при $m \rightarrow \infty$ будет стремиться к квантили нормированного нормального закона распределения порядка $1 - \alpha/2$

$$t_{\alpha, m} \rightarrow Z_{p=1-\alpha/2}.$$

Поэтому, работая с таблицами критерия Стьюдента, неплохо делать проверку, показывающую для какого распределения (одностороннего или двустороннего) они составлены. Так, по табл. П.6

$$t_{0,05;500} = 1,965 \approx Z_{p=1-0,05/2=0,975} = 1,960,$$

следовательно, это двусторонние пределы распределения Стьюдента.

Аналогичная ситуация связана и с функцией **СТЮДРАСПОБР(вероятность;степени_свободы)**, где **вероятность** – это вероятность, соответствующая двустороннему распределению Стьюдента.

Пример 3.6. При проверке Ph-метра с помощью эталонного раствора, имеющего Ph=9,0, получены следующие результаты: 8,7; 9,2; 9,1; 9,0; 9,4; 9,6; 9,7; 8,9; 8,8; 8,7; 9,8; 9,3; 9,8; 8,8, т.е. $n = 14$. Обладает ли Ph-метр систематической погрешностью?

Для решения этой задачи предварительно рассчитаем выборочное среднее \bar{x} и выборочное среднеквадратическое отклонение S в предположении, что показания Ph-метра не противоречат нормальному закону распределения и среди них нет грубых погрешностей (см. формулы (3.5), (3.8) и (3.10)):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{14} \sum_{i=1}^{14} x_i = \frac{1}{14} (8,7 + 9,2 + 9,1 + \dots + 8,8) = 9,2;$$

$$S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] = \frac{1}{14-1} \left[\sum_{i=1}^{14} x_i^2 - \frac{1}{14} \left(\sum_{i=1}^{14} x_i \right)^2 \right] =$$

$$= \frac{1}{14-1} \left[(8,7^2 + 9,2^2 + 9,1^2 + \dots + 8,8^2) - \frac{1}{14} (8,7 + 9,2 + 9,1 + \dots + 8,8)^2 \right] = 0,1646;$$

$$S_x = +\sqrt{S^2} = \sqrt{0,1646} = 0,4057.$$

В электронных таблицах Microsoft Excel для подобных расчетов можно было бы воспользоваться следующими тремя статистическими функциями:

СРЗНАЧ(8,7; 9,2; 9,1; 9; 9,4; 9,6; 9,7; 8,9; 8,8; 8,7; 9,8; 9,3; 9,8; 8,8) = 9,2;

ДИСП(8,7; 9,2; 9,1; 9; 9,4; 9,6; 9,7; 8,9; 8,8; 8,7; 9,8; 9,3; 9,8; 8,8) = 0,164615;

СТАНДОТКЛОН(8,7; 9,2; 9,1; 9; 9,4; 9,6; 9,7; 8,9; 8,8; 8,7; 9,8; 9,3; 9,8; 8,8) = 0,4057.

Далее, в соответствии с описанным выше алгоритмом:

1. Выдвигается нулевая гипотеза о том, что математическое ожидание показаний Ph-метра равно Ph эталонного раствора (не имеют систематической погрешности) $H_0: M_1 = 9$.

2. Альтернативная гипотеза выбирается в виде $H_1: M_1 \neq 9$, поскольку показания Ph-метра не должны как завышать, так и занижать истинное значение Ph раствора.

3. Так как значение генеральной дисперсии σ^2 показаний Ph-метра неизвестно, а имеется только ее оценка $S^2 = 0,1646$, то используется t-критерий (распределения Стьюдента).

4. t - статистика имеет вид (см. (3.49))

$$t = \frac{\bar{x} - M}{S} \cdot \sqrt{n} = \frac{9,2 - 9}{0,4057} \cdot \sqrt{14} = 1,84.$$

5. Выбирается (обычный для большинства технических приложений) уровень значимости $\alpha = 0,05$.

6. При этом уровне значимости, числе степеней свободы $m = n - 1 = 13$ и для альтернативной гипотезы $H_1: M_1 \neq 9$ устанавливаются границы критической

области по табличным значениям квантилей распределения Стьюдента $t_{0,05;13} = 2,16$ или их можно определить, воспользовавшись функцией **СТЮДРАСПОБР**(0,05;13) = 2,160368 из электронных таблиц Microsoft Excel.

7. Поскольку рассчитанное значение статистики $t = 1,84$ не попадает в критическую область ($1,84 < 2,16$), то нулевая гипотеза принимается в качестве рабочей, т.е. можно считать, что $M_1 = 9$ (вероятность того, что показания Рн-метра имеют систематическую погрешность меньше чем 0,05).

В задаче сравнения двух неизвестных математических ожиданий M_1 и M_2 прежде всего рассмотрим ситуацию, когда исследуемые выборки, по которым делаются оценки для M_1 и M_2 , независимы между собой.

Если для двух нормально распределенных генеральных совокупностей с неизвестными параметрами M_1, σ_1^2 и M_2, σ_2^2 получены независимые выборки объемом соответственно n_1 и n_2 , то для сравнения выборочных средних \bar{x}_1 и \bar{x}_2 выдвигается нулевая гипотеза о равенстве математических ожиданий:

1. $H_0: M_1 = M_2$.

2. При этом можно сформулировать три альтернативных гипотезы:

$$H_1^{(1)}: M_1 > M_2; H_1^{(2)}: M_1 < M_2; H_1^{(3)}: M_1 \neq M_2.$$

3. Как и в рассмотренной выше ситуации сравнения с известным математическим ожиданием, используется t-критерий.

4. Вид t-статистики зависит от того, равны $\sigma_1^2 = \sigma_2^2 = \sigma^2$ либо не равны $\sigma_1^2 \neq \sigma_2^2$ между собой генеральные дисперсии (для ответа на этот вопрос можно воспользоваться, например, рассмотренным выше критерием Фишера).

В первом случае, когда дисперсии не имеют значимого отличия, статистика принимает вид

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (3.50)$$

- двухвыборочный t-критерий с равными дисперсиями, где S – обобщенное среднее квадратичное отклонение (см. (3.46)):

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}.$$

Во втором случае, когда дисперсии значительно отличаются друг от друга, $\sigma_1^2 \neq \sigma_2^2$, статистика имеет вид

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad - \quad (3.51)$$

двухвыборочный t-критерий с неравными дисперсиями.

5. В зависимости от условия решаемой задачи выбирается необходимый уровень значимости α .

6. Границы критической области устанавливаются по табличным значениям квантилей t-распределения (см., например, [11] или табл. П.6) либо их можно определить, воспользовавшись статистической функцией **СТЮДРАСПОБР** из электронных таблиц Microsoft Excel. При этом число степеней свободы m рассчитывается:

для $\sigma_1^2 = \sigma_2^2 = \sigma^2$ как $m = n_1 + n_2 - 2$;

для $\sigma_1^2 \neq \sigma_2^2$ $\frac{1}{m} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}$, где $c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

7. Нулевую гипотезу принимают, т.е. полагают, что $M_1 = M_2$ при выполнении неравенств:

- для альтернативных гипотез $H_1^{(1)}$: $M_1 > M_2$; $H_1^{(2)}$: $M_1 < M_2$ $|t| \leq t_{2\alpha, m}$;
- для альтернативной гипотезы $H_1^{(3)}$: $M_1 \neq M_2$ $|t| \leq t_{\alpha, m}$.

Пример 3.7. Проведены испытания механической прочности проб окатышей при использовании старой и двух новых технологий их обжига. Холодная прочность окатышей обычно оценивается при испытании на раздавливание (кН/окатыш). Обычно прочность определяют по результатам раздавливания не менее 20 окатышей размером 12-15 мм.

Для иллюстрации процедуры проверки гипотез о числовых значениях математических ожиданий будем предполагать, что имелась возможность ис-

следовать всего по 8 окатышей для каждой из технологий. Результаты испытаний представлены в табл. 3.4.

Таблица 3.4

*Результаты испытаний прочности окатышей,
изготовленных по разным технологиям, кН/окатыш*

Номер окатыша	Старая технология X_{0i}	Новая технология, вариант 1 X_{1i}	Новая технология, вариант 2 X_{2i}	$X_{\Delta i} = X_{1i} - X_{2i}$
1	2,11	2,21	2,21	0
2	2,12	2,26	2,22	0,04
3	1,97	2,19	2,08	0,11
4	2,10	2,21	2,19	0,02
5	2,17	2,27	2,24	0,03
6	2,12	2,24	2,21	0,03
7	1,93	2,14	2,06	0,08
8	2,28	2,32	2,31	0,01
\bar{x}	2,10	2,23	2,19	0,04
S^2	0,0120	0,003029	0,0068	0,001371

Можно ли по полученным данным сделать вывод, что новая технология по варианту 1 позволяет повысить прочность окатышей?

1. Сформулируем нулевую гипотезу $H_0: M_1 = M_0$.

2. Поскольку предполагается, что новая технология по варианту 1 позволит повысить прочность окатышей, то альтернативная гипотеза выбирается в виде $H_1: M_1 > M_0$.

3. Будем считать, что выборки взяты из генеральных совокупностей с нормальным законом распределения. Для того чтобы определить вариант статистики для t – критерия, сравним между собой соответствующие дисперсии. Для этого в качестве нулевой гипотезы примем $H_0: \sigma_1^2 = \sigma_0^2 = \sigma^2$. В предположении, что новая технология позволяет также снизить и разброс в значениях прочности (т.е. иметь и более стабильный технологический процесс), в качестве альтернативной гипотезы примем $H_1: \sigma_1^2 < \sigma_0^2$.

Статистика F -критерия (критерия Фишера) при этом равна $F=0,0120/0,003029 = 3,96$, и для построения критической области при $\alpha = 0,05$ находим $F_{0,05;8-1;8-1} = 3,79$ (по таблицам либо в Microsoft Excel через **FPAC-**

ПОБР(0,05;7;7) = 3,787051). Поскольку $3,96 > 3,79$, то с вероятностью большей чем 0,95 можно говорить, что $\sigma_1^2 < \sigma_0^2$.

4. t - статистика в этом случае должна иметь вид (см. (3.51))

$$t = \frac{2,23 - 2,10}{\sqrt{\frac{0,003029}{8} + \frac{0,0120}{8}}} = 3,00.$$

5. Как обычно, выберем уровень значимости $\alpha = 0,05$.

6. Для построения критической области рассчитаем число степеней свободы:

$$c = \frac{\frac{0,03029}{8}}{\frac{0,03029}{8} + \frac{0,0120}{8}} = 0,20; \quad \frac{1}{m} = \frac{0,20^2}{8-1} + \frac{(1-0,20)^2}{8-1} = 0,096883; \quad m = 10,3.$$

Табличное значение $t_{2*0,05;10} = 1,81$ (**СТЮДРАСПОБР**(0,1;10) = 1,812462).

7. Поскольку рассчитанное ранее значение статистики попадает в критическую область $3,00 > 1,81$, то нулевая гипотеза $H_0: M_1 = M_0$ должна быть отвергнута, т.е. новая технология по варианту 1 действительно позволяет повысить прочность окатышей.

Вероятность ошибки подобного утверждения (ошибки первого рода, заключающейся в том, что отвергают нулевую гипотезу $H_0: M_1 = M_0$, в то время как в действительности эта гипотеза верна), т.е. уровень значимости α при этом можно оценить как **СТЮДРАСП**(3,00;10;1) = 0,006672. При расчете значения функции распределения Стьюдента в данном случае используется: найденная в пункте 4 статистика $t = 3,00$; определенное в пункте 6 число степеней свободы $m \approx 10$ и такой параметр, как число возвращаемых "хвостов" распределения. "Хвосты" = 1, и функция **СТЮДРАСП** возвращает одностороннее распределение, поскольку была принята односторонняя альтернативная гипотеза $H_1: M_1 > M_0$.

Для определения найденного выше значения уровня значимости $\alpha = 0,0067$ в электронных таблицах в Microsoft Excel может быть использована такая статистическая функция, как **ТТЕСТ**. Она используется для того, чтобы оп-

ределить, насколько вероятно, что две выборки взяты из генеральных совокупностей, которые имеют одно и то же математическое ожидание:

ТТЕСТ({2,21;2,26;2,19;2,21;2,27;2,24;2,14;2,32};{2,11;2,12;1,97;2,10;2,17;2,12;1,93;2,28};1;3)=0,006459.

В качестве аргументов функции **ТТЕСТ**, кроме самих выборочных значений (которые стоят в фигурных скобках), используется еще такие два параметра, как "Хвосты" = 1 (для односторонней альтернативной гипотезы) и "Тип" - это вид исполняемого t-теста. В данном случае "Тип" = 3, поскольку необходимо провести двухвыборочный t-тест с неравными дисперсиями. Полученное в данном случае значение t-теста говорит о том, что вероятность равенства математического ожидания прочности окатышей по новой (вариант 1) и старой технологии их обжига очень мала (составляет только 0,6%), следовательно, новая технология по варианту 1 действительно является более предпочтительной, чем старая.

В продолжение примера 3.7 ответим на вопрос: есть или нет какое-либо значимое различие между двумя новыми технологиями обжига (по варианту 1 и 2) с точки зрения повышения механической прочности окатышей?

1. В соответствии с общим алгоритмом проверки статистических гипотез сформулируем $H_0: M_1 = M_2$.

2. Поскольку предполагается, что обе новые технологии равнозначны между собой, то альтернативная гипотеза выбирается в виде $H_1: M_1 \neq M_2$.

3. Для того чтобы определить тип t – теста, сравним между собой дисперсии $H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2$ в предположении, что обе новые технологии дают одинаковый разброс в значениях прочности, альтернативная гипотеза выбирается в виде $H_1: \sigma_1^2 \neq \sigma_2^2$.

Статистика критерия Фишера при этом равна $F = 0,0068/0,003029 = 2,25$ (в числителе критерия Фишера всегда должна стоять большая дисперсия), а поскольку при $\alpha = 0,05$ $F_{(0,05/2);8-1;8-1} = 4,99$ (см. табл. П.5) и $F < F_{(0,05/2);8-1;8-1}$ ($2,25 < 4,99$), то действительно можно считать, что $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

S – обобщенное среднее квадратичное отклонение тогда будет равно (см.(3.46))

$$S = \sqrt{\frac{(8-1)0,003029 + (8-1)0,0068}{8+8-2}} = 0,07.$$

4. Поскольку дисперсии не имеют значимого отличия, статистика t – критерия принимает вид (см. (3.50))

$$t = \frac{2,23 - 2,19}{0,07 \sqrt{\frac{1}{8} + \frac{1}{8}}} = 1,14.$$

5. Выбираем уровень значимости $\alpha = 0,05$ и определяем число степеней свободы $m = 8 + 8 - 2 = 14$.

6. Для построения критической области находим табличное значение $t_{0,05;14} = 2,15$ (**СТЮДРАСП**(0,05;14) = 2,144789).

7. Поскольку рассчитанное ранее значение статистики не попадает в критическую область $1,14 < 2,15$, то нулевая гипотеза $H_0: M_1 = M_2$ принимается в качестве рабочей, т.е. новые технологии как по варианту 1, так и по варианту 2 равнозначны между собой с точки зрения повышения механической прочности окатышей.

Вероятность ошибки (первого рода) при этом можно оценить величиной **СТЮДРАСП**(1,14;14;2) = 0,272934, т.е. если бы мы в подобных ситуациях отвергали нулевую гипотезу, то примерно в 27 случаях из 100 мы поступали неверно. В данном случае "Хвосты" = 2, и функция **СТЮДРАСП** возвращает двустороннее распределение, поскольку альтернативная гипотеза была принята в виде $H_1: M_1 \neq M_2$, а не в виде $H_1: M_1 > M_2$.

Для определения найденного нами значения уровня значимости $\alpha = 0,27$ в электронных таблицах Microsoft Excel также могла быть использована функция

$$\mathbf{TTEST}(\{2,21;2,26;2,19;2,21;2,27;2,24;2,14;2,32\};\{2,21;2,22;2,08;2,19;2,24;2,21;2,06;2,31\};2;2) = 0,272934.$$

В данном случае "Хвосты" = 2 (для двусторонней альтернативной гипотезы) и "Тип" = 2, поскольку используется двухвыборочный t -тест с равными дисперсиями.

И наконец, в задаче сравнения двух неизвестных математических ожиданий M_1 и M_2 рассмотрим ситуацию, когда исследуемые выборки зависимы между собой.

t-критерий для зависимых выборок очень полезен в тех довольно часто возникающих на практике ситуациях, когда важный источник внутригрупповой вариации (разброса или ошибки) может быть легко определен и исключен из анализа. Это относится к экспериментам, в которых две сравниваемые группы получены на одной и той же совокупности наблюдений (субъектов), которые тестировались дважды (например, до и после термообработки проката, до и после вакуумирования стали, измерения, производимые на одних и тех же партиях продукции различными методами или различными приборами и т.д.). В подобных экспериментах значительная часть внутригрупповой изменчивости (вариации) в обеих группах может быть объяснена индивидуальными различиями субъектов (различиями в свойствах отдельных прокатанных полос, каждой конкретной плавки или партии продукции). Если та же самая выборка тестируется дважды, то можно легко исключить эту часть вариации. Вместо исследования каждой группы отдельно можно рассматривать просто разности между двумя измерениями для каждого субъекта (например, анализировать одни и те же плавки "до вакуумирования" и "после вакуумирования"). Вычитая первые значения из вторых (для каждого субъекта: прокатанной полосы, плавки или партии продукции) и анализируя затем только эти "чистые (парные) разности", появляется возможность исключить ту часть вариации, которая является результатом различия в исходных уровнях индивидуумов. Именно так и проводятся вычисления в t-критерии для зависимых выборок. В сравнении с t-критерием для независимых выборок такой подход дает всегда "лучший" результат (критерий становится более чувствительным).

Реализация t-критерия для зависимых выборок начинается с того, что строится новая выборка из $n = n_1 = n_2$ элементов (парные наблюдения), определяемая как разность значений первой и второй выборок: $x_{\Delta i} = x_{1i} - x_{2i}$ и по ней рассчитываются оценки математического ожидания \bar{x}_{Δ} и среднеквадратичного отклонения S_{Δ} :

$$\bar{x}_{\Delta} = \frac{1}{n} \cdot \sum_{i=1}^n x_{\Delta i}; S_{\Delta} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (\bar{x}_{\Delta} - x_{\Delta i})^2}.$$

1. Выдвигается нулевая гипотеза о том, что математическое ожидание разности равно нулю $H_0: M_{\Delta} = 0$.

2. Как и для случая независимых выборок, можно сформулировать три альтернативных гипотезы:

$$H_1^{(1)}: M_{\Delta} > 0; H_1^{(2)}: M_{\Delta} < 0; H_1^{(3)}: M_{\Delta} \neq 0;$$

3. Используется t-критерий для зависимых выборок (парный).

4. Статистика критерия Стьюдента, учитывая, что $M_{\Delta} = 0$, примет вид (см. (3.49))

$$t = \frac{\bar{x}_{\Delta} - M_{\Delta}}{S_{\Delta}} \cdot \sqrt{n} = \frac{\bar{x}_{\Delta} - 0}{S_{\Delta}} \cdot \sqrt{n} = \frac{\bar{x}_{\Delta}}{S_{\Delta}} \cdot \sqrt{n}. \quad (3.52)$$

5. В зависимости от условия решаемой задачи выбирается необходимый уровень значимости α . Число степеней свободы для зависимых выборок равно $m = n - 1$.

6. Границы критической области устанавливаются в зависимости от вида альтернативной гипотезы по значениям квантилей распределения Стьюдента $t_{\alpha; m}$ или $t_{2\alpha; m}$.

7. Нулевую гипотезу принимают, т.е. полагают, что $M_{\Delta} = 0$ при выполнении неравенств:

- для альтернативных гипотез $H_1^{(1)}: M_{\Delta} > 0; H_1^{(2)}: M_{\Delta} < 0$ $|t| \leq t_{2\alpha; m};$

- для альтернативной гипотезы $H_1^{(3)}: M_{\Delta} \neq 0$

$$|t| \leq t_{\alpha; m}.$$

Еще раз обратимся к числовому материалу примера 3.7 и переформулируем условия задачи таким образом, чтобы как по варианту 1, так и по варианту 2 были приведены данные для одной и той же новой технологии, полученные дважды на одних и тех же партиях окатышей, но измерения прочности выполнены по двум различным методикам.

Можно ли сказать, что результаты измерения прочности, полученные для новой технологии по различным методикам на одних и тех же партиях окатышей, не имеют значимого различия?

Поскольку при таких условиях задачи выборки по варианту 1 и 2 становятся зависимыми друг от друга (значения прочности окатышей по каждой из восьми партий произведены дважды, но по разным методикам), то для решения необходимо воспользоваться описанным выше парным t-критерием.

Рассчитанные значения $x_{\Delta i}$, \bar{x}_{Δ} и S_{Δ}^2 приведены в табл. 3.4 (см. последний столбец).

$$S_{\Delta} = \sqrt{0,001371} = 0,037 .$$

1. Выдвигаем нулевую гипотезу $H_0: M_{\Delta} = 0$.

2. Поскольку между двумя методиками не предполагается никакого различия, то альтернативную гипотезу выбираем в виде $H_1: M_{\Delta} \neq 0$

3. Используется t-критерий для зависимых выборок (парный).

4. Статистика критерия Стьюдента в этом случае представляет собой

$$t = \frac{\bar{x}_{\Delta}}{S_{\Delta}} \cdot \sqrt{n} = \frac{0,04}{0,037} \cdot \sqrt{8} = 3,055.$$

5. Выбираем уровень значимости $\alpha = 0,05$ и определяем число степеней свободы $m = 8 - 1 = 7$.

6. Для построения критической области находим табличное значение $t_{0,05;7} = 2,37$ (**СТЮДРАСПОБР**(0,05;7) = 2,364623).

7. Поскольку рассчитанное ранее значение статистики попадает в критическую область $3,06 > 2,37$, то нулевая гипотеза $H_0: M_1 = 0$ отвергается, и в качестве рабочей необходимо принять альтернативную $H_1: M_{\Delta} \neq 0$, т.е. методики определения прочности по варианту 1 и по варианту 2 дают значимо различные результаты на одних и тех же партиях и для одной и той же новой технологии отжига окатышей.

Вероятность ошибки первого рода при этом составляет ("Хвосты" = 2, поскольку $H_1: M_1 \neq 0$) **СТЮДРАСП**(3,055;7;2) = 0,018453, т.е., отвергая в анало-

гичных условиях нулевую гипотезу, мы примерно только в одном или двух случаях из 100 будем допускать ошибку.

Найденная оценка $\alpha = 0,018$ в электронных таблицах Microsoft Excel может быть рассчитана с использованием функции **ТТЕСТ**.

ТТЕСТ({2,21;2,26;2,19;2,21;2,27;2,24;2,14;2,32};{2,21;2,22;2,08;2,19;2,24;2,21;2,06;2,31};2;1)= 0,018452.

Последний параметр в этой функции "Тип" = 1 (парный t-тест).

Если сравнить результаты, полученные в примере 3.7 по t-тесту на двух совершенно одинаковых выборках (вариант 1 и 2) при условии, что эти выборки независимы (двухвыборочный t-тест с равными дисперсиями) и зависимы (парный t-тест), то можно увидеть, что они дают совершенно противоположные результаты. Когда на выборки по варианту 1 и 2 мы смотрели как на независимые, мы не видели различия в их математических ожиданиях, но при условии зависимости в математических ожиданиях удалось установить значимые расхождения. Этот числовой материал подтверждает ранее уже высказанное положение о том, что t-критерий для зависимых выборок является более чувствительным.

Поскольку методика парного t-теста полностью повторяет алгоритм сравнения неизвестного математического ожидания M_1 с конкретным числовым значением M , то статистическая функция **ТТЕСТ** в электронных таблицах Microsoft Excel применима и для решения задач о соответствии полученного в эксперименте выборочного среднего \bar{x}_1 известному математическому ожиданию.

Так, для примера 3.6 (о наличии погрешности в показаниях Ph-метра). **ТТЕСТ**({8,7;9,2;9,1;9,9;9,4;9,6;9,7;8,9;8,8;8,7;9,8;9,3;9,8;8,8};{9;9;9;9;9;9;9;9;9;9;9;9;9;9;9};2;1) = 0,088025, что при найденном в этом примере значении статистики $t = 1,84$, числе степеней свободы $m = 14 - 1 = 13$ и альтернативной гипотезе $H_1: M_1 \neq 9$ ("Хвосты" = 2) соответствует **СТЮДРАСП**(1,84;13;2) = 0,088706. Полученное значение функции **ТТЕСТ** говорит о том, что вероятность наличия

систематической погрешности у Ph-метра может быть оценена величиной $1 - 0,089 = 0,91$ (меньшей, чем 0,95, значения которого мы закладывали, выбирая уровень значимости $\alpha = 0,05$).

В заключение этого раздела еще раз подчеркнем, что все перечисленные выше критерии могут быть использованы только для случайных величин, не противоречащих нормальному закону распределения (закону распределения Гаусса). Так, например, применительно к t-критерию для зависимых выборок это означает, что попарные разности должны быть нормально распределены. Если это предположение не выполняется (о том, как его можно проверить, смотри следующий раздел), то необходимо воспользоваться одним из альтернативных непараметрических критериев (см. например, [10]).

3.6. Критерии согласия. Проверка гипотез о виде функции распределения

Рассмотренные ранее методы оценивания параметров распределения случайной величины и критерии для проверки статистических гипотез предполагали, что известна функция распределения (нормальный закон – распределение Гаусса). Однако в большинстве случаев вид закона распределения является гипотетическим и сам по себе требует статистического подтверждения.

Наиболее простым, но весьма приближенным методом проверки согласия результатов эксперимента с тем или иным законом распределения является графический метод. Он заключается в оценке эмпирической функции распределения и сопоставлении ее с функцией предполагаемого теоретического закона. Если построенные экспериментальные точки лежат вблизи теоретического графика, то можно считать, что полученные в опытах данные не противоречат выбранному теоретическому закону распределения. Графический метод является в значительной мере субъективным и используется на практике в качестве первого приближения при решении подобных задач.

Более объективные методы установления вида распределения случайной величины строятся на аппарате проверки статистических гипотез - критериях согласия.

Нулевая гипотеза в данном случае заключается в том, что H_0 : исследуемая генеральная совокупность не противоречит предполагаемому теоретическому закону распределения. При этом альтернативная гипотеза обычно формулируется как H_1 : случайная величина имеет любое другое распределение, отличное от предполагаемого.

Разработано достаточно много критериев согласия, отличающихся как своей мощностью, так и объемом опытных данных, необходимых для их использования. Рассмотрим некоторые из них, и в первую очередь остановимся на критериях согласия, которые могут быть использованы при относительно больших объемах выборки.

Когда экспериментатор располагает достаточно представительным количеством экспериментальных данных ($n > 100$), то их предварительная обработка начинается с группировки, которая проводится в следующей последовательности:

1. Находят наибольшее (x_{\max}) и наименьшее (x_{\min}) выборочные значения случайной величины и вычисляют ее размах $R = x_{\max} - x_{\min}$.

2. Размах случайной величины разбивают на k равных интервалов. Количество интервалов k выбирают в зависимости от объема выборки. Например, при $n > 100$ его значение рекомендуется принимать равным $k = 9 \div 15$ (при $n < 100$ $k = 7$). Число интервалов k можно определить и по формуле Штюргеса $k = 1 + 3,32 \lg(n)$ с округлением полученного значения до ближайшей целой величины.

3. Определяют ширину интервала $h = R/k$, для упрощения расчетов полученные значения округляют в любую сторону, несколько увеличивая или уменьшая при этом размах варьирования R .

4. Устанавливают границы интервалов и подсчитывают число попаданий случайной величины в каждый из выбранных интервалов \bar{m}_i , $1 \leq i \leq k$.

5. Определяют частоту попаданий для каждого интервала как $P_i = \bar{m}_i / n$. Результаты подобных вычислений могут быть сведены в таблицу (подобную, например, табл. 3.5).

Таблица 3.5

Построение распределения экспериментальных данных

Интервал	Число замеров в каждом интервале \bar{m}_i	Частота попадания в интервал $P_i = \bar{m}_i / n$
$X_1 \div X_2$	\bar{m}_1	\bar{m}_1 / n
$X_2 \div X_3$	\bar{m}_2	\bar{m}_2 / n
...
$X_i \div X_{i+1}$	\bar{m}_i	\bar{m}_i / n
...
$X_k \div X_{k+1}$	\bar{m}_k	\bar{m}_k / n
Проверка	$\sum_{i=1}^k \bar{m}_i = n$	$\sum_{i=1}^k P_i = 1$

Графической формой представления непрерывной случайной величины является гистограмма (рис.3.8). Последовательность построения гистограмм следующая:

1. Определяется величина ординаты $f_i = P_i / h$, где P_i – вероятность появления случайной величины в i -м интервале.

2. В системе координат $f_i = f(x)$ на ширине интервала h откладывают величины f_i как высоты и строятся прямоугольники.

Очевидно, что площадь элементарного прямоугольника

$$S_i = h \cdot f_i = h \cdot \frac{P_i}{h} = P_i = \frac{\bar{m}_i}{n} \quad (3.53)$$

равна отношению числа опытов m_i , при которых случайная величина оказалась внутри этого интервала, к общему числу опытов n .

Площадь всей гистограммы $S = \sum_{i=1}^k S_i = \sum_{i=1}^k P_i = 1$. Следовательно, площадь, ограниченная гистограммой, равна единице.

3. Построение гистограммы интегральной функции распределения осуществляется суммированием вероятностей: $F(x) = \sum_{i=1}^k P_i$.

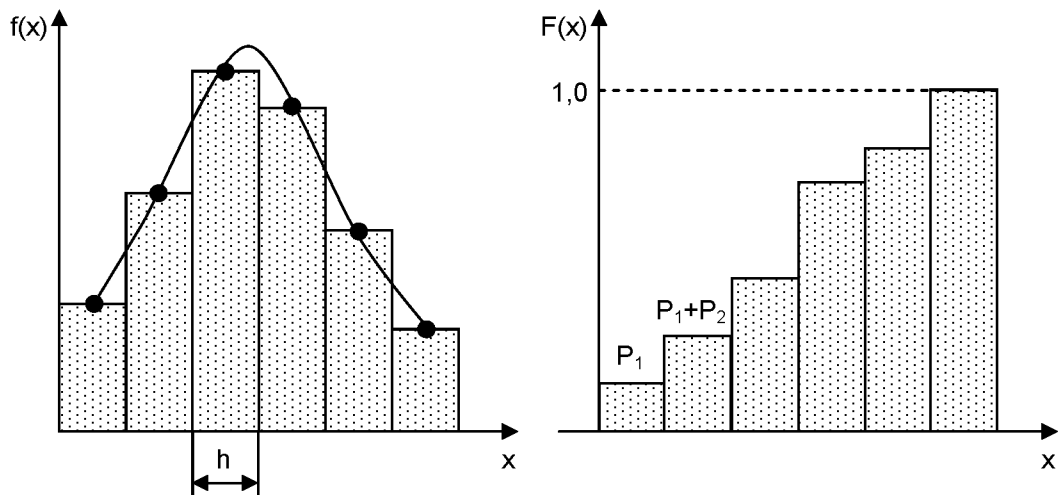


Рис.3.8. К построению гистограммы случайной величины

В дальнейшем осуществляется сравнение экспериментально полученного распределения случайной величины с некоторым видом теоретического распределения. Для этой цели используются различные критерии согласия: χ^2 (хи-квадрат) Пирсона, Колмогорова–Смирнова и др.

Критерий Пирсона

Рассмотрим методику проверки гипотезы нормального распределения по критерию χ^2 Пирсона. Этот критерий кроме определения доверительного интервала для дисперсии нередко используется для проверки согласованности распределений, полученных по данным выборки с некоторой теоретической плотностью распределения.

В данном случае применение критерия χ^2 предполагает использование свойств нормированного (стандартного) нормального распределения. Напомним, что уравнение кривой плотности стандартного нормального распределения имеет вид

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \approx 0,4 \cdot e^{-z^2/2}; \quad z = \frac{x - M_x}{\sigma_x}.$$

Тогда теоретическая вероятность попадания случайной величины в интервал $\Delta z = z_{i+1} - z_i$ в случае нормального распределения можно определить по формуле

$$P_i^* = F(z_{i+1}) - F(z_i) = \frac{1}{2\pi} \int_{z_i}^{z_{i+1}} e^{-u^2/2} du. \quad (3.54)$$

Отличие оценки закона распределения P от теоретического закона распределения P^* можно охарактеризовать величиной

$$\chi^2 = \sum_{i=1}^k C_i (P_i - P_i^*)^2, \quad (3.55)$$

где P_i и P_i^* – оценка и теоретическая вероятность случайной величины для i -го интервала; C_i – весовые коэффициенты, которые с большим весом учитывают отклонения для меньших P_i .

Пирсон выбрал весовые коэффициенты следующим образом:

$$C_i = \frac{n}{P_i^*}. \quad (3.56)$$

Пирсон показал, что при таком выборе C_i закон распределения χ^2 слабо зависит от n и $P(x)$, а определяется в основном числом разрядов k .

Следовательно,

$$\chi^2 = n \sum_{i=1}^k \frac{(P_i - P_i^*)^2}{P_i^*} = n \sum_{i=1}^k \frac{(\bar{m}_i/n - P_i^*)^2}{P_i^*} = \sum_{i=1}^k \frac{(\bar{m}_i - n \cdot P_i^*)^2}{n \cdot P_i^*}. \quad (3.57)$$

Очевидно, что при идеальном соответствии экспериментальных данных нормальному закону распределения экспериментальное значение критерия Пирсона будет равно нулю, т.к. $P_i = P_i^*$.

В выражении (3.55) стоит сумма квадратов k случайных величин, однако они не являются независимыми, так как на них накладывается некоторое число связей. Одной из таких связей является требование, чтобы площадь под кривой

оценки закона распределения равнялась единице: $\sum_{i=1}^k P_i = 1$. Иногда требуют,

чтобы среднее значение \bar{x} совпадало с математическим ожиданием M_x , а выборочная дисперсия S_x^2 – с дисперсией σ_x^2 . Поэтому число степеней свободы чаще всего определяется как

$$m = k - 2. \quad (3.58)$$

Теоретическое значение критерия Пирсона $\chi^2_{\alpha;m}$ определяется по справочным данным (см. табл.П.3) или с использованием пакетов прикладных программ при заданном уровне значимости α и числе степеней свободы m (см. функцию **ХИ2ОБР**($\alpha;m$) из электронных таблиц Microsoft Excel).

Алгоритм использования критерия Пирсона заключается в следующем.

1. Выдвигаются нуль-гипотеза H_0 : "Отличие экспериментальных данных от нормального закона распределения не существенно" и альтернативная ей гипотеза H_1 : "Отличие экспериментальных данных от нормального закона распределения существенно, т.е. экспериментальные данные не подчиняются закону нормального распределения".

2. По результатам экспериментальных измерений и предположению нормального закона их распределения определяется расчетное значение критерия Пирсона χ^2 .

3. Определяют число степеней свободы m , задаются уровнем значимости α и определяют теоретическое значение критерия Пирсона $\chi^2_{\alpha;m}$.

4. Если $\chi^2 < \chi^2_{\alpha;m}$, то нуль-гипотеза H_0 о нормальном законе распределения экспериментальных данных принимается с доверительной вероятностью $P=1-\alpha$. В противном случае нуль-гипотеза отвергается и принимается альтернативная гипотеза H_1 .

Отметим важные рекомендации по использованию критерия χ^2 .

Если при некотором числе измерений критерий $\chi^2 > \chi^2_{\alpha;m}$, но сомнения в нормальности распределения отсутствуют, то следует, если имеется возможность, увеличить число измерений в несколько раз и повторить анализ по этому же критерию.

Число степеней свободы $m=k-2$ относится к такому случаю, когда оба параметра нормального закона распределения определяются по результатам измерений, т.е. когда вместо точных измерений значений M_x и σ_x применяют их эмпирические значения (оценки) \bar{x} и S_x . Если же значение M_x точно известно (например, при измерении эталона), то число степеней свободы равно $k=n-1$; если известны оба параметра M_x и σ_x , то число степеней свободы равно $k=n$. На практике такая ситуация встречается относительно редко, и поэтому для получения числа степеней свободы не менее пяти желательно брать число интервалов не менее семи (иногда девяти).

Критерий Колмогорова–Смирнова

Рассмотрим использование критерия Колмогорова–Смирнова для проверки гипотезы нормальности распределения случайных величин. Данная процедура также предполагает построение таблицы распределения экспериментально полученных значений с группировкой данных в определенное число разрядов k . Дополнительно в таблицу необходимо включить следующие колонки:

- колонку с накопительной суммой $\sum_{p=1}^i \bar{m}_p$ для каждого i -го интервала

сгруппированных данных, как показано в табл.3.6.

Таблица 3.6

Процедура вычисления критерия Колмогорова–Смирнова

Интервал	Число замеров в каждом интервале \bar{m}_i	$\sum_{p=1}^i \bar{m}_p$	Теоретическая вероятность P_i^*	$n \sum_{p=1}^i P_p^*$	$\left \sum_{p=1}^i \bar{m}_p - n \sum_{p=1}^i P_p^* \right $
$X_1 \div X_2$	\bar{m}_1	\bar{m}_1	P_1^*	nP_1^*	$ \bar{m}_1 - nP_1^* $
$X_2 \div X_3$	\bar{m}_2	$\bar{m}_1 + \bar{m}_2$	P_2^*	$n(P_1^* + P_2^*)$	$ (\bar{m}_1 + \bar{m}_2) - n(P_1^* + P_2^*) $
...
$X_i \div X_{i+1}$	\bar{m}_i	$\sum_{p=1}^i \bar{m}_p$	P_i^*	$n \sum_{p=1}^i P_p^*$	$\left \sum_{p=1}^i \bar{m}_p - n \sum_{p=1}^i P_p^* \right $
...

$X_k \div X_{k+1}$	\bar{m}_k	$\sum_{p=1}^k \bar{m}_p$	P_k^*	$n \sum_{p=1}^k P_p^*$	$\left \sum_{p=1}^k \bar{m}_p - n \sum_{p=1}^k P_p^* \right $
--------------------	-------------	--------------------------	---------	------------------------	--

– колонку с накопительной суммой $n \sum_{p=1}^i P_p^*$, где P^* – теоретическая

вероятность попадания случайной величины в i -й интервал, значения которой, как уже отмечалось ранее, табулированы и приводятся в статистических справочниках.

На основании данных табл.3.6 вычисляют экспериментальное значение критерия согласия Колмогорова–Смирнова:

$$D = \frac{\max \left| \sum_{p=1}^i \bar{m}_p - n \sum_{p=1}^i P_p^* \right|}{n}. \quad (3.58a)$$

Далее экспериментальное значение критерия Колмогорова–Смирнова сравнивают с теоретическим $D_{n;\alpha}$, которое определяют из статистических таблиц в соответствии с объемом выборки n и требуемым уровнем значимости α (см. табл. П.10). Если $D < D_{n;\alpha}$, то гипотеза о нормальном распределении результатов замеров принимается с вероятностью $P=1-\alpha$.

Критерий согласия Колмогорова–Смирнова для проверки нормальности распределения результатов наблюдений входит в пакет статистической обработки данных STATISTICA, с которым мы познакомимся в главе 7.

Рассмотрим использование рассмотренных критериев χ^2 Пирсона и Колмогорова–Смирнова для оценки нормальности распределения данных на следующем примере.

Пример 3.8. В табл. 3.7 приведено содержание кремния в чугуна при выплавке передельного чугуна в доменной печи, которое изменяется в пределах от 0,32 до 0,95%. Всего было отобрано 50 проб чугуна.

Требуется оценить, подчиняется ли содержание кремния в пробах нормальному закону распределения? Если да, определить медиану, моду, среднее, выборочную дисперсию, доверительный интервал для математического ожидания, среднеквадратичное отклонение. Для вычислений

воспользуемся статистическими функциями из электронной таблицы Microsoft Excel.

Предварительно вычислим с использованием статистических функций **СРЗНАЧ**, **ДИСП** и **СТАНДОТКЛОН** среднее значение \bar{x} , выборочную дисперсию S_x^2 и стандартное отклонение S_x , которые оказались равны соответственно $\bar{x}=0,65$, $S_x^2=0,01853$ и $S_x=0,1361$.

Таблица 3.7

Содержание кремния в чугуна по результатам отбора 50 проб

Номер пробы	1	2	3	4	5	6	7	8	9	10
[Si],%	0,32	0,35	0,45	0,43	0,41	0,51	0,52	0,53	0,57	0,58
Номер пробы	11	12	13	14	15	16	17	18	19	20
[Si],%	0,59	0,56	0,56	0,58	0,54	0,57	0,61	0,62	0,63	0,64
Номер пробы	21	22	23	24	25	26	27	28	29	30
[Si],%	0,65	0,66	0,67	0,68	0,69	0,61	0,65	0,62	0,63	0,67
Номер пробы	31	32	33	34	35	36	37	38	39	40
[Si],%	0,65	0,62	0,68	0,71	0,72	0,78	0,75	0,72	0,79	0,72
Номер пробы	41	42	43	44	45	46	47	48	49	50
[Si],%	0,73	0,72	0,79	0,73	0,84	0,82	0,87	0,90	0,95	0,93

Примем число интервалов равным 7. Тогда величина интервала составит $h=(0,95-0,32)/7=0,09=0,1$. Результаты группировки исходных данных и вычислений приведены в табл. 3.8.

Таблица 3.8

Процедура вычисления критерия χ^2 Пирсона по данным примера 3.8

Интервал $x_{i-1} \div x_i$	\bar{m}_i	$F(x_i)$	$P_i^* = F(x_i) - F(x_{i-1})$	nP_i^*	$\bar{m}_i - nP_i^*$	$\chi_i^2 = \frac{(\bar{m}_i - nP_i^*)^2}{nP_i^*}$
0,3÷0,4	2	0,033	0,033	1,7	0,4	0,07
0,4÷0,5	3	0,135	0,102	5,1	-2,1	0,87
0,5÷0,6	11	0,356	0,221	11,1	-0,1	0,00
0,6÷0,7	17	0,642	0,286	14,3	2,7	0,51
0,7÷0,8	11	0,864	0,222	11,1	-0,1	0,00

0,8÷0,9	4	0,967	0,103	5,2	-1,2	0,26
0,9÷1,0	2	0,995	0,028	1,4	0,6	0,26

Вычисление $F(x_i)$ проводили с использованием статистической функции **НОРМРАСП**. В частности, для интервала 0,3÷0,4 находим

НОРМРАСП(0,4; СРЗНАЧ(B4:B53); СТАНДОТКЛОН(B4:B53);ИСТИНА)=0,033.

Отметим, что поскольку среди аргументов функции **НОРМРАСП** есть среднее арифметическое и стандартное отклонение, то для определения соответствующих параметров также воспользуемся встроенными функциями электронных таблиц Microsoft Excel **СРЗНАЧ()** и **СТАНДОТКЛОН()**. В показанном примере полагаем, что данные 50 опытов по содержанию кремния в чугуне расположены на листе электронной таблицы в ячейках от B4 до B53. Аналогично определяли функции распределения для каждого интервала, результаты отражены в табл. 3.8.

Таким образом, экспериментальное значение критерия Пирсона $\chi^2 = \sum_{i=1}^k \chi_i^2 = 1,96$, а теоретическое при уровне значимости $\alpha=0,05$ и числе степеней свободы $m_1=7-2=5$ составляет $\chi^2_{0,05;5}=11,07$ (**ХИ2ОБР(0,05;5)=11,07048**), что значительно больше экспериментального значения.

Следовательно, весьма уверенно можно утверждать, что содержание кремния в пробах чугуна подчиняется нормальному закону распределения.

Предлагаем читателям самостоятельно оценить доверительный интервал математического ожидания. Здесь можно отметить, что, как показывают расчеты, доверительный интервал с вероятностью 95% равен 0,030. Учитывая близость распределения к нормальному и достаточно большое число экспериментальных точек, можно воспользоваться электронными таблицами Microsoft Excel (функция **ДОВЕРИТ**), т.е. действительное среднее содержание кремния в чугуне (математическое ожидание) лежит в интервале от 0,62% до 0,68%.

Процедура проверки гипотезы нормального распределения данных из примера с использованием критерия Колмогорова–Смирнова (D) представлена в табл.3.9.

Таблица 3.9

Процедура вычисления критерия Колмогорова–Смирнова
по данным примера 3.8

Интервал $x_{i-1} \div x_i$	\bar{m}_i	$\sum_{p=1}^i \bar{m}_p$	$F(x_i)$	$P_i^* = F(x_i) - F(x_{i-1})$	nP_i^*	$n \sum_{p=1}^i P_p^*$	$\left \sum_{p=1}^i \bar{m}_p - n \sum_{p=1}^i P_p^* \right $
0,3÷0,4	2	2	0,033	0,033	1,7	1,7	0,4
0,4÷0,5	3	5	0,135	0,102	5,1	6,8	1,8
0,5÷0,6	11	16	0,356	0,221	11,1	17,8	1,8
0,6÷0,7	17	33	0,642	0,286	14,3	32,1	0,9
0,7÷0,8	11	44	0,864	0,222	11,1	43,2	0,8
0,8÷0,9	4	48	0,967	0,103	5,2	48,4	0,3
0,9÷1,0	2	50	0,995	0,028	1,4	49,8	0,3

На основании результатов этой таблицы определяем максимальное значение из последней колонки и по нему рассчитываем экспериментальное значение критерия D:

$$D = \frac{\max \left| \sum_{p=1}^i \bar{m}_p - n \sum_{p=1}^i P_p^* \right|}{n} = \frac{1,8}{50} = 0,036.$$

Экспериментальное значение критерия сравниваем с теоретическим $D_{50;0,05} = 0,177$, взятым из табл. П.10 с учетом объема выборки $n=50$ и уровнем значимости $\alpha=0,05$. Величина $D_{50;0,05} > D$, поэтому можно сделать тот же вывод, что и ранее: гипотеза нормального распределения результатов измерения содержания кремния в чугуна принимается с вероятностью 95%.

Пример 3.9. В табл. 3.10 приведено содержание оксида железа в шлаке перед раскислением металла при выплавке стали марки 0,8 КП в 200 -тонной мартеновской печи, работающей с применением кислорода для интенсификации горения топлива и прямого окисления примесей ванны. Всего отобрано 56 проб шлака.

Таблица 3.10

Исходные данные к примеру 3.9

$x_i, \%$:	\overline{m}_i	$\sum_{i=1}^1 \overline{m}_i$
10	1	1
12	5	6
13	3	9
14	6	15
15	6	21
16	2	23
17	8	31
18	6	37
19	6	43
20	4	47
22	1	48
23	4	52
24	1	53
25	1	54
26	1	55
30	1	56
31	0	56

$$\overline{x} = \frac{\sum_{i=1}^{17} x_i \cdot m_i^*}{\sum_{i=1}^{17} m_i} = 17,375;$$

$$S_x^2 = \frac{\sum_{i=1}^{17} m_i \cdot (x_i - \overline{x})^2}{\sum_{i=1}^{17} m_i - 1} = 16,0;$$

$$S_x = 4,0.$$

Из табл. 3.10 видно, что содержание оксида железа в шлаке колеблется от 10 до 31%. Примем число интервалов равным $k=7$. Тогда величина интервала составит $h=(31-10)/7=3$. Результаты группировки исходных данных и вычислений приведены в табл. 3.11. Расчетное значение критерия Пирсона составило $\chi^2=3,861$. Его следует сравнить с табличным значением $\chi_{\alpha;m}^2$ для 5%-ного уровня значимости, которое при числе степеней свободы, равном $m=5$, составляет $\chi_{0,05;5}^2=11,07$.

Таблица 3.11

Процедура вычисления критерия χ^2 по данным примера 3.9

Интервал $X_{i-1} \div X_i$	\bar{m}_i	$F(x_i)$	$P_i = F(x_i) - F(x_{i-1})$	nP_i	$m_i - nP_i$	$(m_i - nP_i)$	$\chi_i^2 = \frac{(m_i - nP_i)^2}{nP_i}$
10÷13	6	0,0326 0,137	0,104	5,824	0,176	0,031	0,005
13÷16	15	0,366	0,229	12,824	2,176	4,735	0,369
16÷19	16	0,658	0,292	16,352	0,352	0,124	0,008
19÷22	10	0,876	0,218	12,208	2,208	4,875	0,399
22÷25	6	0,972	0,096	5,376	0,624	0,389	0,072
25÷28	2	0,996	0,024	1,344	0,656	0,430	0,320
28÷31	1	1,000	0,004	0,224	0,776	0,602	2,688
Сумма	56	-	-	-	-	-	$\Sigma=3,861$

Таким образом, опытные данные не противоречат гипотезе о том, что содержание FeO в конечном шлаке подчиняется закону нормального распределения. Оцените самостоятельно моду, медиану, доверительный интервал для математического ожидания при вероятностях 90, 95 и 99,7%.

3.7. Преобразование распределений к нормальному

Если исследователь, используя методы, изложенные в предыдущем параграфе, убедился, что гипотеза нормальности распределения не может быть принята, то вполне может быть, что с помощью существующих методов удастся так преобразовать исходные данные, что их распределение будет подчиняться нормальному закону распределения. Для пояснения идеи преобразований рассмотрим качественный пример. Пусть кривая распределения $f(x)$ имеет вид, представленный на рис. 3.9,а, т.е. имеются очень крутая левая ветвь и пологая правая. Такое распределение отличается от нормального.

Для выполнения операций преобразования каждое наблюдение трансформируется с помощью логарифмического преобразования $x' = \lg(x)$. При этом левая ветвь кривой распределения сильно растягивается, и распределение принимает приближенно нормальный вид (рис.3.9,б). Если при

преобразовании получают значения, расположенные между 0 и 1, то все наблюдаемые значения для удобства расчетов и во избежание получения отрицательных параметров необходимо умножить на 10 в соответствующей степени, чтобы все вновь полученные, преобразованные значения были больше единицы, т.е. необходимо выполнить преобразования $x'' = \lg(x \cdot 10^a)$.

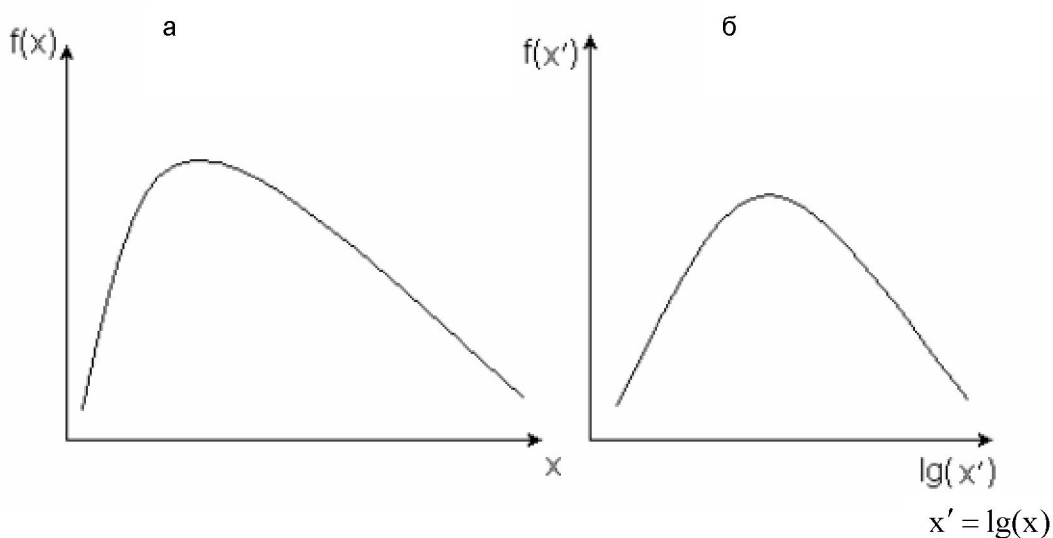


Рис. 3.9. Преобразование функции $f(x)$ к нормальному распределению

Асимметричное распределение с одной вершиной приводится к нормальному преобразованием $x' = \lg(x \pm a)$. В отдельных случаях можно применять и другие преобразования:

- а) обратная величина $x' = 1/x$;
- б) обратное значение квадратных корней $x' = 1/\sqrt{x}$.

Преобразование "обратная величина" является наиболее "сильным". Среднее положение между логарифмическим преобразованием и "обратной величиной" занимает преобразование "обратное значение квадратных корней".

Для нормализации смещенного вправо распределения служат, например, степенные преобразования $x' = x^a$. При этом для a принимают значения: $a=1,5$ при умеренном и $a=2$ при сильно выраженном правом смещении. Рекомендуем читателю придумать такие преобразования, которые удовлетворяли бы исследователя в том или ином случае.

4. АНАЛИЗ РЕЗУЛЬТАТОВ ПАССИВНОГО ЭКСПЕРИМЕНТА. ЭМПИРИЧЕСКИЕ ЗАВИСИМОСТИ

4.1. Характеристика видов связей между рядами наблюдений

На практике сама необходимость измерений большинства величин вызывается тем, что они не остаются постоянными, а изменяются в функции от изменения других величин. В этом случае целью проведения эксперимента является установление вида функциональной зависимости $\hat{y}=f(\mathbf{X})$. Для этого должны одновременно определяться как значения \mathbf{X} , так и соответствующие им значения \hat{y} , а задачей эксперимента является установление математической модели исследуемой зависимости. Фактически речь идет об установлении связи между двумя рядами наблюдений (измерений).

Определение связи включает в себя указание вида модели и определение ее параметров. В теории экспериментов независимые параметры $\mathbf{X}=(x_1, \dots, x_k)$ принято называть факторами, а зависимые переменные y – откликами. Координатное пространство с координатами $x_1, x_2, \dots, x_i, \dots, x_k$ называется факторным пространством. Эксперимент по определению вида функции

$$\hat{y} = f(x), \quad (4.1)$$

где x – скаляр, называется однофакторным. Эксперимент по определению функции вида

$$\hat{y} = f(\mathbf{X}), \quad (4.1a)$$

где $\mathbf{X}=(x_1, x_2, \dots, x_i, \dots, x_k)$ – вектор – многофакторным.

Геометрическим представлением функции отклика в факторном пространстве является поверхность отклика. При однофакторном эксперименте ($k=1$) поверхность отклика представляет собой линию на плоскости, при двухфакторном ($k=2$) – поверхность в трехмерном пространстве.

Связи в общем случае являются достаточно многообразными и сложными. Обычно выделяют следующие виды связей.

Функциональные связи (или зависимости) – это такие связи, когда при изменении величины X другая величина y изменяется так, что каждому значению x_i соответствует совершенно определенное (однозначное) значение y_i (рис.4.1,а). Таким образом, если выбрать все условия эксперимента абсолютно одинаковыми, то, повторяя испытания, получим одну и ту же зависимость, т.е. кривые идеально совпадут для всех испытаний.

К сожалению, такие условия в реальности не встречаются. На практике не удастся поддерживать постоянство условий (например, физико-химические свойства шихты при моделировании процессов тепломассопереноса в металлургических печах). При этом влияние каждого случайного фактора в отдельности может быть мало, однако в совокупности они существенно могут повлиять на результаты эксперимента. В этом случае говорят о стохастической (вероятностной) связи между переменными.

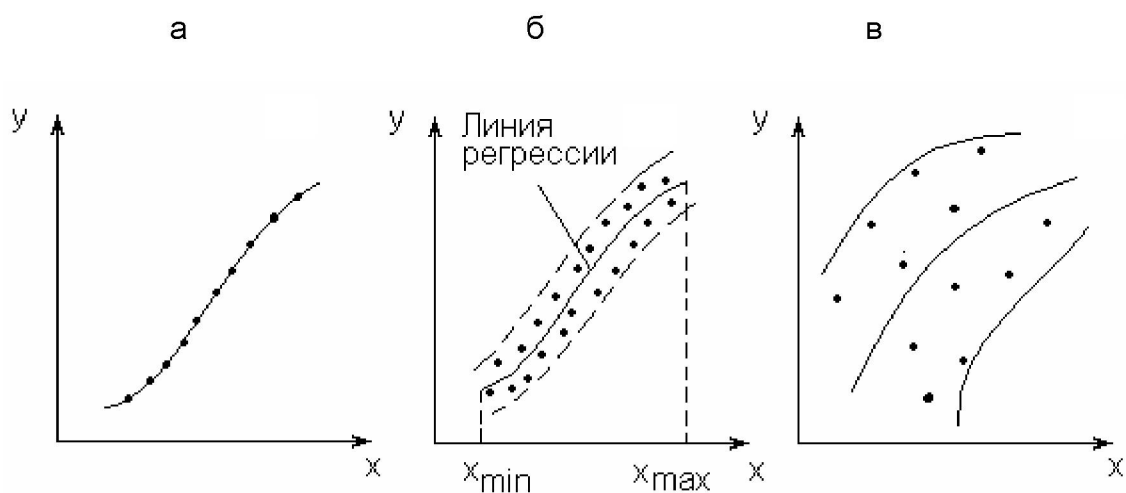


Рис.4.1. Виды связей: а – функциональная связь, все точки лежат на линии; б – связь достаточно тесная, точки группируются возле линии регрессии, но не все они лежат на ней; в – связь слабая

Стохастичность связи состоит в том, что одна случайная переменная y реагирует на изменение другой X изменением своего закона распределения (см. рис. 4.1, б). Таким образом, зависимая переменная принимает не одно конкретное значение, а некоторое из множества значений. Повторяя испытания, мы будем получать другие значения функции отклика, и одному и тому же значению X в различных реализациях будут соответствовать различные значения

у в интервале $[x_{\min}; x_{\max}]$. Искомая зависимость $\hat{y} = f(\mathbf{X})$ может быть найдена лишь в результате совместной обработки полученных значений \mathbf{X} и y .

На рис. 4.1, б – это кривая зависимости, проходящая по центру полосы экспериментальных точек (математическому ожиданию), которые могут и не лежать на искомой кривой $\hat{y} = f(\mathbf{X})$, а занимают некоторую полосу вокруг нее. Эти отклонения вызваны погрешностями измерений, неполнотой модели и учитываемых факторов, случайным характером самих исследуемых процессов и другими причинами.

Анализ стохастических связей приводит к различным постановкам задач статистического исследования зависимостей, которые упрощенно можно классифицировать следующим образом:

- 1) задачи корреляционного анализа – задачи исследования наличия взаимосвязей между отдельными группами переменных ;
- 2) задачи регрессионного анализа – задачи, связанные с установлением аналитических зависимостей между переменным y и одним или несколькими переменными $x_1, x_2, \dots, x_i, \dots, x_k$, которые носят количественный характер;
- 3) задачи дисперсионного анализа – задачи, в которых переменные $x_1, x_2, \dots, x_i, \dots, x_k$ имеют качественный характер, а исследуется и устанавливается степень их влияния на переменное y .

Стохастические зависимости характеризуются формой, теснотой связи и численными значениями коэффициентов уравнения регрессии.

Форма связи устанавливает вид функциональной зависимости $\hat{y} = f(\mathbf{X})$ и характеризуется уравнением регрессии. Если уравнение связи линейное, то имеем линейную многомерную регрессию, в этом случае зависимость \hat{y} от \mathbf{X} описывается линейной зависимостью в k -мерном пространстве:

$$\hat{y} = b_0 + \sum_{j=1}^k b_j x_j, \quad (4.2)$$

где $b_0, \dots, b_j, \dots, b_k$ – коэффициенты уравнения. Для пояснения существа используемых методов ограничимся сначала случаем, когда x – скаляр. В общем

случае виды функциональных зависимостей в технике достаточно многообразны: показательные $y = b_0 x^{b_1}$, логарифмические $y = b_0 \lg(x)$ и т.д.

Заметим, что задача выбора вида функциональной зависимости – задача неформализуемая, так как одна и та же кривая на данном участке примерно с одинаковой точностью может быть описана самыми различными аналитическими выражениями. Отсюда следует важный практический вывод. Даже в наш век компьютеров принятие решения о выборе той или иной математической модели остается за исследователем. Только экспериментатор знает, для чего будет в дальнейшем использоваться эта модель, на основе каких понятий будут интерпретироваться ее параметры.

Крайне желательно при обработке результатов эксперимента вид функции $\hat{y}=f(X)$ выбирать, исходя из условия ее соответствия физической природе изучаемых явлений или имеющимся представлениям об особенностях поведения исследуемой величины. К сожалению, такая возможность не всегда имеется, так как эксперименты чаще всего проводятся для исследования недостаточно или неполно изученных явлений.

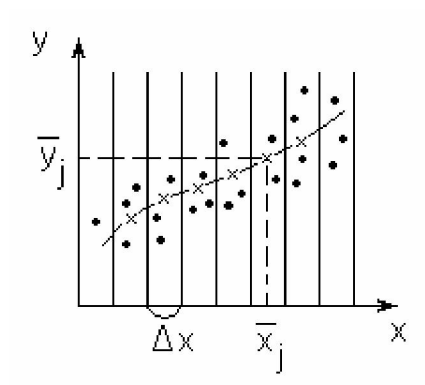


Рис.4.2. К построению эмпирической линии регрессии

При изучении зависимости $\hat{y}=f(x)$ от одного фактора при заранее неизвестном виде функции отклика для приближенного определения вида уравнения регрессии полезно предварительно построить эмпирическую линию регрессии (рис.4.2). Для этого весь диапазон изменения x разбивают на равные интервалы Δx . Все точки, попавшие в данный интервал Δx_j , относят к его середине \bar{x}_j . Для этого подсчитывают частные

средние для каждого интервала:

$$\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{ji}}{n_j}. \quad (4.3)$$

Здесь n_j – число точек в интервале Δx_j , причем $\sum_{j=1}^{k^*} n_j = n$, где k^* – число интервалов разбиения; n – объем выборки.

Затем последовательно соединяют точки $(\overline{x}_j; \overline{y}_j)$ отрезками прямой. Полученная ломаная называется эмпирической линией регрессии. По виду эмпирической линии регрессии можно в первом приближении подобрать вид уравнения регрессии $\hat{y} = f(x)$.

Под теснотой связи понимается степень близости стохастической зависимости к функциональной, т.е. показатель тесноты группирования экспериментальных данных относительно принятого уравнения модели (см. рис. 4.1,б,в). В дальнейшем уточним это положение.

4.2. Определение коэффициентов уравнения регрессии

Будем полагать, что вид уравнения регрессии уже выбран и требуется определить только конкретные численные значения коэффициентов этого уравнения $\mathbf{b} = \{b_0, \dots, b_j, \dots, b_k\}$. Отметим предварительно, что если выбор вида уравнения регрессии, как это уже отмечалось, – процесс неформальный и не может быть полностью передан компьютеру, то расчет коэффициентов выбранного уравнения регрессии – операция достаточно формальная и ее следует решать с использованием компьютера. Это трудный и утомительный расчет, в котором человек не застрахован от ошибок, а компьютер выполнит его значительно быстрее и качественнее.

Существует два основных подхода к нахождению коэффициентов b_j . Выбор того или иного из них определяется целями и задачами, стоящими перед исследователем, точностью полученных результатов, их количеством и т.д.

Первый подход – интерполирование. Базируется на удовлетворении условию, чтобы функция $\hat{y} = (\mathbf{X}, \mathbf{b})$ совпадала с экспериментальными значениями в некоторых точках, выбранных в качестве опорных (основных, главных) y_i .

В этом случае для определения $k+1$ неизвестных значений параметров b_j используется система уравнений

$$f(x_i, b_0, \dots, b_j, \dots, b_k) = y_i, \quad 1 \leq i \leq n. \quad (4.4)$$

В данном случае число независимых уравнений системы равно числу опорных точек, в пределе – n поставленных опытов. С другой стороны, для определения $k+1$ коэффициентов необходимо не менее $k+1$ независимых уравнений. Но если число n поставленных опытов и число независимых уравнений равно числу искомых коэффициентов $k+1$, то решение системы может быть единственно, а следовательно, точно соответствует случайным значениям исходных данных. Таким образом, в предельном случае, когда число коэффициентов уравнения регрессии равно

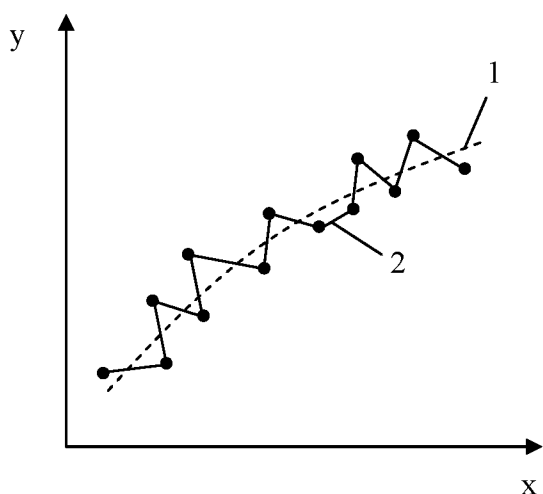


Рис.4.3. Аппроксимация функции с большим (1) и небольшим (2) числом коэффициентов b_i

числу экспериментальных точек $n=k+1$, все экспериментальные точки будут совпадать с их расчетными значениями. Следует заметить, что добиваться такого точного совпадения путем значительного увеличения числа коэффициентов уравнения регрессии часто просто неразумно, поскольку экспериментальные результаты получены с большей или меньшей погрешностью, и

такая функция может просто не отражать действительного характера изменения исследуемой величины в силу влияния помех (возмущений) (рис.4.3).

Таким образом, задача в конечном счете сводится к решению системы $k+1$ уравнений с $k+1$ неизвестными. Основная сложность такого решения связана с нелинейностью системы, хотя в принципе при использовании компьютера она преодолима.

При числе опытов n большем, чем $k+1$ искомым коэффициентов, число независимых уравнений системы избыточно. Избыточность информации можно использовать по-разному.

После определения численных значений $k+1$ параметров проверяется качество аппроксимации путем сопоставления значений функции и экспериментальных данных в оставшихся, неиспользованных точках. Если обнаруженные между ними расхождения превышают допустимые по условию точности, то процедуру определения коэффициентов b_j можно повторить, приняв в качестве опорных (основных) другие точки.

Таким образом, из этих уравнений в разных комбинациях можно составить несколько систем уравнений, каждая из которых в отдельности даст свое решение. Но между собой они будут несовместимыми. Каждое решение будет соответствовать своим значениям коэффициентов b_j . Если все их построить на графике, то получим целый пучок аппроксимирующих кривых.

Это открывает при $n > k+1$ совершенно новые возможности. Во-первых, этот пучок кривых показывает форму и ширину области неопределенности проведенного эксперимента. Во-вторых, может быть произведено усреднение всех найденных кривых и полученная усредненная кривая будет гораздо точнее и достовернее описывать исследуемое явление, так как она в значительной степени освобождена от случайных погрешностей, приводивших к разбросу отдельных экспериментальных точек. Поясним суть этого подхода на примере двух методов.

1. Метод избранных точек (рис. 4.4). На основании анализа данных выдвигают гипотезу о виде (форме) зависимости $f(\mathbf{X})$. Предположим, что она линейная, т.е. статистическая связь – это линейная одномерная регрессия

$$\hat{y} = b_0 + b_1 x. \quad (4.5)$$

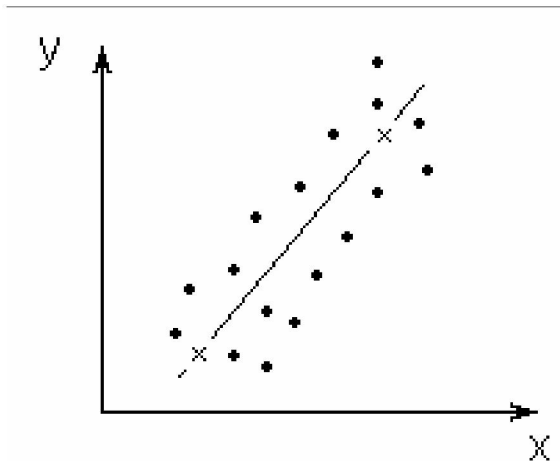


Рис.4.4. Метод избранных точек:
x – избранные точки

Выбирают две наиболее характерные, по мнению исследователя, точки, через которые и проходит линия регрессии (рис. 4.4). Задача вычисления коэффициентов b_0 и b_1 в этом случае тривиальна. Если предполагается, что уравнение регрессии более высокого порядка, то соответственно увеличивают число избранных точек. Недостатки такого подхода очевидны, так как избранные точки выбираются субъективно, а подавляющая часть экспериментального материала не используется для определения параметров (коэффициентов) уравнения регрессии, хотя ее можно использовать в дальнейшем для оценки надежности полученного уравнения.

2. Метод медианных центров. Сущность этого метода поясняет рис.4.5. Обведенное контуром поле точек делят на несколько частей, число которых равно числу определяемых коэффициентов уравнения регрессии. В каждой из этих частей находят медианный центр, т.е. пересечение вертикали и горизонтали слева и справа, выше и ниже которых оказывается равное число точек. Затем через эти медианные центры проводят плавную кривую и из решения системы уравнений определяют коэффициенты регрессии b_j .

Так, в случае линейной зависимости (4.5) поле делится на две группы. Определяют средние значения $\bar{x}_I, \bar{y}_I; \bar{x}_{II}, \bar{y}_{II}$ для каждой из групп, а неизвестные коэффициенты b_0, b_1 определяют из решения системы уравнений:

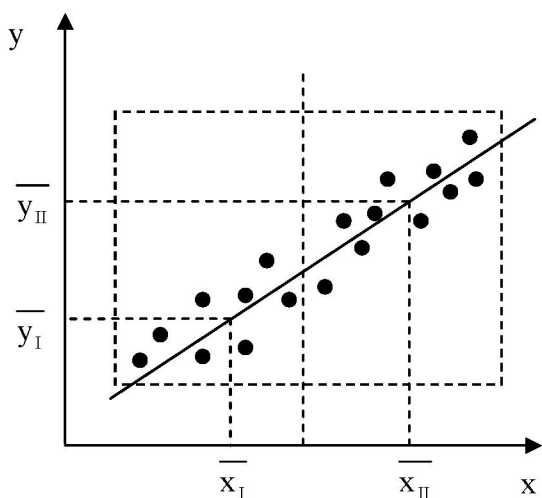


Рис.4.5. Метод медианных точек

Обведенное контуром поле точек делят на несколько частей, число которых равно числу определяемых коэффициентов уравнения регрессии. В каждой из этих частей находят медианный центр, т.е. пересечение вертикали и горизонтали слева и справа, выше и ниже которых оказывается равное число точек. Затем через эти медианные центры проводят плавную кривую и из решения системы уравнений определяют коэффициенты регрессии b_j .

Так, в случае линейной зависимости (4.5) поле делится на две группы. Определяют средние значения $\bar{x}_I, \bar{y}_I; \bar{x}_{II}, \bar{y}_{II}$ для каждой из групп, а неизвестные коэффициенты b_0, b_1 определяют из решения системы уравнений:

$$\begin{aligned}\overline{y_I} &= b_0 + b_1 \overline{x_I}; \\ \overline{y_{II}} &= b_0 + b_1 \overline{x_{II}}.\end{aligned}\tag{4.5a}$$

Если при выборе вида уравнения регрессии число его коэффициентов b_j окажется больше числа уравнений (имеющихся результатов измерений) $k+1 > n$, система (4.4) не будет иметь однозначного решения. В этом случае необходимо либо уменьшить число определяемых коэффициентов $k+1$, либо увеличить число опытов n .

Второй подход – метод наименьших квадратов. Усреднение несовместимых решений избыточной системы уравнений $n > k+1$ может быть преодолено методом наименьших квадратов, который был разработан еще Лежандром и Гауссом. Таким образом, метод наименьших квадратов – это «новинка» почти 200-летней давности. Сегодня, благодаря возможностям компьютеров, этот метод вступил, по существу, в полосу своего «ренессанса».

Определение коэффициентов b_j методом наименьших квадратов основано на выполнении требования, чтобы сумма квадратов отклонений экспериментальных точек от соответствующих значений уравнения регрессии была минимальна. Заметим, что, в принципе, можно оперировать и суммой других четных степеней этих отклонений, но тогда вычисления будут сложнее. Однако руководствоваться суммой отклонений нельзя, так как она может оказаться малой при больших отклонениях отрицательного знака.

Математическая запись приведенного выше требования имеет вид

$$\Phi(b_0, b_1, \dots, b_j, \dots, b_k) = \sum_{i=1}^n [f(x_i, b_0, b_1, \dots, b_j, \dots, b_k) - y_i]^2 \rightarrow \min_{b_j}, \tag{4.6}$$

где n – число экспериментальных точек в рассматриваемом интервале изменения аргумента x .

Необходимым условием минимума функции $\Phi(b_0, b_1, \dots, b_j, \dots, b_k)$ является выполнение равенства

$$\partial \Phi / \partial b_j = 0, \quad 0 \leq j \leq k \tag{4.7}$$

или

$$\sum_{i=1}^n [f(x_i, b_0, b_1, \dots, b_j, \dots, b_k) - y_i] \frac{\partial f(x_i)}{\partial b_j} = 0, \quad 0 \leq j \leq k. \quad (4.7a)$$

После преобразований получим

$$\sum_{i=1}^n [f(x_i, b_0, b_1, \dots, b_j, \dots, b_k) \frac{\partial f(x_i)}{\partial b_j} - \sum_{i=1}^n y_i \frac{\partial f(x_i)}{\partial b_j}] = 0. \quad (4.8)$$

Система уравнений (4.8) содержит столько же уравнений, сколько неизвестных коэффициентов b_0, b_1, \dots, b_k входит в уравнение регрессии, и называется в математической статистике системой нормальных уравнений.

Поскольку $\Phi \geq 0$ при любых b_0, \dots, b_k , величина Φ обязательно должна иметь хотя бы один минимум. Поэтому если система нормальных уравнений имеет единственное решение, оно и является минимумом для этой величины.

Расчет регрессионных коэффициентов методом наименьших квадратов можно применять при любых статистических данных, распределенных по любому закону.

4.3. Определение тесноты связи между случайными величинами

Определив уравнение теоретической линии регрессии, необходимо дать количественную оценку тесноты связи между двумя рядами наблюдений. Линии регрессии, проведенные на рис. 4.1, б, в, одинаковы, однако на рис. 4.1, б точки значительно ближе (теснее) расположены к линии регрессии, чем на рис. 4.1, в.

При корреляционном анализе предполагается, что факторы и отклики носят случайный характер и подчиняются нормальному закону распределения.

Тесноту связи между случайными величинами характеризуют корреляционным отношением ρ_{xy} . Остановимся подробнее на физическом смысле данного показателя. Для этого введем новые понятия.

Остаточная дисперсия $S_{y_{ост}}^2$ характеризует разброс экспериментально наблюдаемых точек относительно линии регрессии и представляет собой показатель ошибки предсказания параметра y по уравнению регрессии (рис. 4.6):

$$S_{y_{\text{ост}}}^2 = \frac{1}{n-l} \sum_{i=1}^n [y_i - \hat{y}_i]^2 = \frac{1}{n-1-k} \sum_{i=1}^n [y_i - f(x_i, b_0, b_1, \dots, b_k)]^2, \quad (4.9)$$

где $l=k+1$ – число коэффициентов уравнения модели.

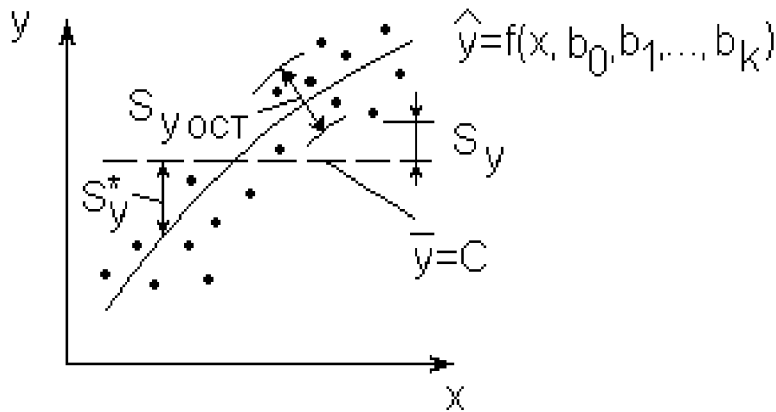


Рис.4.6. К определению дисперсий

Общая дисперсия (дисперсия выходного параметра) S_y^2 характеризует разброс экспериментально наблюдаемых точек относительно среднего значения \bar{y} , т.е. линии C (см. рис. 4.6):

$$S_y^2 = \frac{1}{n-1} \sum_{i=1}^n [y_i - \bar{y}]^2, \quad (4.10)$$

где $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Средний квадрат отклонения линии регрессии от среднего значения линии $\bar{y} = C$ (см. рис. 4.6):

$$S_y^{*2} = \frac{1}{n-1} \sum_{i=1}^n [\hat{y}_i - \bar{y}]^2 = \frac{1}{n-1} \sum_{i=1}^n [f(x_i, b_0, b_1, \dots, b_k) - \bar{y}]^2. \quad (4.11)$$

Очевидно, что общая дисперсия S_y^2 (сумма квадратов относительно среднего значения \bar{y}) равна остаточной дисперсии $S_{y_{\text{ост}}}^2$ (сумме квадратов относительно линии регрессии) плюс средний квадрат отклонения линии регрессии S_y^{*2} (сумма квадратов, обусловленная регрессией).

$$S_y^2 = S_{y \text{ ост}}^2 + S_y^{*2}. \quad (4.11a)$$

Разброс экспериментально наблюдаемых точек относительно линии регрессии характеризуется безразмерной величиной – выборочным корреляционным отношением, которое определяет долю, которую привносит величина X в общую изменчивость случайной величины Y.

$$\rho_{xy}^* = \sqrt{\frac{S_y^2 - S_{y \text{ ост}}^2}{S_y^2}} = \sqrt{\frac{S_y^{*2}}{S_y^2}} = \frac{S_y^*}{S_y}. \quad (4.12)$$

Проанализируем свойства этого показателя.

1. В том случае, когда связь является не стохастической, а функциональной, корреляционное отношение равно 1, так как все точки корреляционного поля оказываются на линии регрессии, остаточная дисперсия равна $S_{y \text{ ост}}^2 = 0$, а $S_y^{*2} = S_y^2$ (рис. 4.7, а).

2. Равенство нулю корреляционного отношения указывает на отсутствие какой-либо тесноты связи между величинами x и y для данного уравнения регрессии, поскольку разброс экспериментальных точек относительно среднего значения и линии регрессии одинаков, т.е. $S_y^2 = S_{y \text{ ост}}^2$ (рис. 4.7, б).

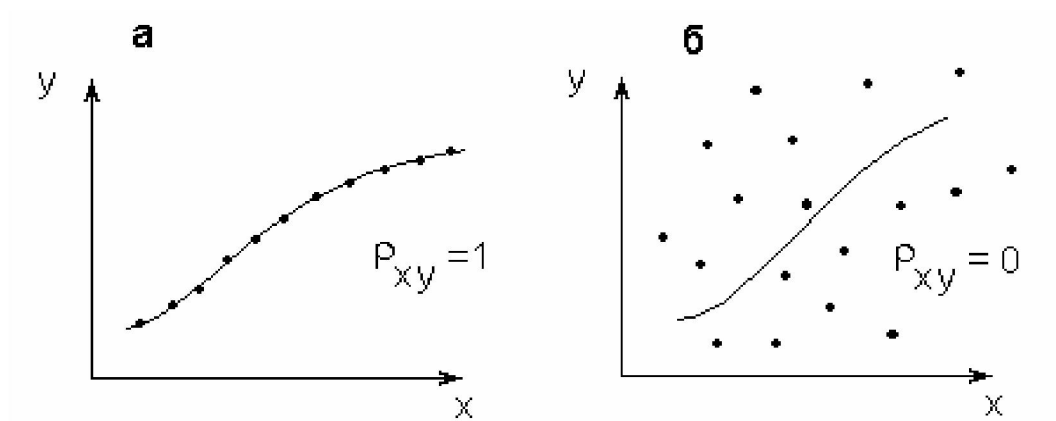


Рис. 4.7. Значения выборочного корреляционного отношения ρ_{xy} :

a – функциональная связь; б – отсутствие связи

3. Чем ближе расположены экспериментальные данные к линии регрессии, тем теснее связь, тем меньше остаточная дисперсия и тем больше корреляционное отношение.

Следовательно, корреляционное отношение может изменяться в пределах от 0 до 1.

Учитывая, что для компьютеров имеются пакеты программ для статистической обработки результатов исследований, рассмотрим методологию этого подхода на примере простейших линейных и одномерных задач (см. уравнение (4.5)). Идеология решения более сложных задач принципиально не отличается. Более того, как мы увидим в дальнейшем, многие нелинейные зависимости можно свести к линейным.

4.4. Линейная регрессия от одного фактора

Уравнение линии регрессии на плоскости в декартовых координатах имеет вид выражения (4.5).

Задачу метода наименьших квадратов аналитически можно выразить следующим образом:

$$\Phi(b_0, b_1) = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2 \rightarrow \min_{b_0, b_1} . \quad (4.13)$$

Для решения этой задачи, как известно из математического анализа, необходимо вычислить частные производные функции Φ по коэффициентам b_0 , b_1 и приравнять их нулю:

$$\frac{\partial \Phi(b_0, b_1)}{\partial b_0} = 0; \quad \frac{\partial \Phi(b_0, b_1)}{\partial b_1} = 0. \quad (4.14)$$

Система нормальных уравнений (4.8) в этом случае примет вид

$$\begin{cases} \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)] = 0; \quad n b_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \\ \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)] \cdot x_i = 0; \quad b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{cases} \quad (4.15)$$

Решение этой системы относительно b_0 и b_1 дает

$$b_0 = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n (x_i)^2 - \sum_{i=1}^n (x_i y_i) \sum_{i=1}^n x_i}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2}; \quad (4.16)$$

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (4.16a)$$

т.е. для расчета b_0 и b_1 необходимо определить $\sum x_i, \sum y_i, \sum x_i y_i, \sum (x_i)^2$.

Коэффициент b_0 (свободный член уравнения регрессии) геометрически представляет собой расстояние от начала координат до точки пересечения линии регрессии с осью ординат, а коэффициент b_1 характеризует тангенс угла наклона линии регрессии к оси Ox .

Если же определяют уравнение регрессии в виде

$$\hat{y} = b_0 + b_1 x + b_{11} x^2,$$

то система уравнений для нахождения b_0, b_1, b_{11} будет иметь следующий вид:

$$\begin{cases} \sum_{i=1}^n y_i = b_0 n + b_1 \sum_{i=1}^n x_i + b_{11} \sum_{i=1}^n x_i^2, \\ \sum_{i=1}^n x_i y_i = b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 + b_{11} \sum_{i=1}^n x_i^3, \\ \sum_{i=1}^n x_i^2 y_i = b_0 \sum_{i=1}^n x_i^2 + b_1 \sum_{i=1}^n x_i^3 + b_{11} \sum_{i=1}^n x_i^4. \end{cases} \quad (4.16б)$$

Из уравнений (4.15) и (4.16б) вытекает правило записи любых систем нормальных уравнений: необходимо записать столько уравнений в системе,

сколько неизвестных коэффициентов содержится в искомом уравнении, всякий раз суммируя произведения членов исходного уравнения на переменную при искомом коэффициенте.

Оценку силы линейной связи осуществляют по выборочному (эмпирическому) коэффициенту парной корреляции r_{xy} . Выборочный коэффициент корреляции может быть вычислен двумя способами.

1. Как частный случай корреляционного отношения для линейного уравнения регрессии.

С учетом того, что $\bar{y} = b_0 + b_1 \bar{x}$,

$$S_y^{*2} = \frac{1}{n-1} \sum_{i=1}^n [b_0 + b_1 x_i - b_0 - b_1 \bar{x}]^2 = b_1^2 S_x^2, \quad (4.17)$$

величина отношения S_y^*/S_y будет равна

$$r_{xy} = b_1 S_x / S_y, \quad (4.18)$$

где S_x и S_y – выборочные средние квадратичные отклонения.

2. Как среднее значение произведения центрированных случайных величин, отнесенное к произведению их среднеквадратичных отклонений:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (4.19)$$

Покажем, что две последние формулы эквивалентны. Для этого преобразуем выражение (4.19) к виду

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

Подставляя последнее выражение в формулу (4.16а), имеем

$$b_1 = \frac{r_{xy} (n-1) S_x S_y}{\sum_{i=1}^n (x_i - \bar{x})^2} = r_{xy} S_y / S_x, \text{ откуда } r_{xy} = \frac{b_1 S_x}{S_y}.$$

Как правило, по результатам экспериментов находят S_x , S_y , \bar{x} , \bar{y} и рассчитывают r_{xy} по формуле (4.19), а затем, используя эти величины, определяют коэффициенты уравнения регрессии:

$$b_1 = r_{xy} S_y / S_x; \quad b_0 = \bar{y} - b_1 \bar{x}. \quad (4.20)$$

Коэффициент корреляции r_{xy} изменяется в пределах $-1 \leq r_{xy} \leq +1$.

Положительная корреляция между случайными величинами характеризует такую стохастическую зависимость между величинами, когда с возрастанием одной из них другая в среднем также будет возрастать. При отрицательной корреляции с возрастанием одной случайной величины другая в среднем будет уменьшаться. Чем ближе значение r_{xy} к единице, тем теснее статистическая связь.

Отметим еще раз область применимости выборочного коэффициента корреляции для оценки тесноты связи.

1. Коэффициент парной корреляции значений y и x применительно к однофакторной зависимости характеризует тесноту группирования данных лишь относительно прямой (например, линия А на рис. 4.8, а). При более сложной зависимости (рис. 4.8, б) коэффициент корреляции r_{xy} будет оценивать тесноту экспериментальных точек относительно некоторой прямой, обозначенной буквой А, что, естественно, несет мало сведений о тесноте их группирования относительно искомой кривой $\hat{y} = f(x)$.
2. Коэффициент парной выборочной корреляции имеет четкий физический смысл только в случае двумерного нормального распределения параметров, т.е. когда для каждого значения X , например x_1 , x_2 , x_3 , существует совокупность нормального распределения y и наоборот, а дисперсия зависимой переменной при изменении значения аргумента остается постоянной (рис. 4.9).

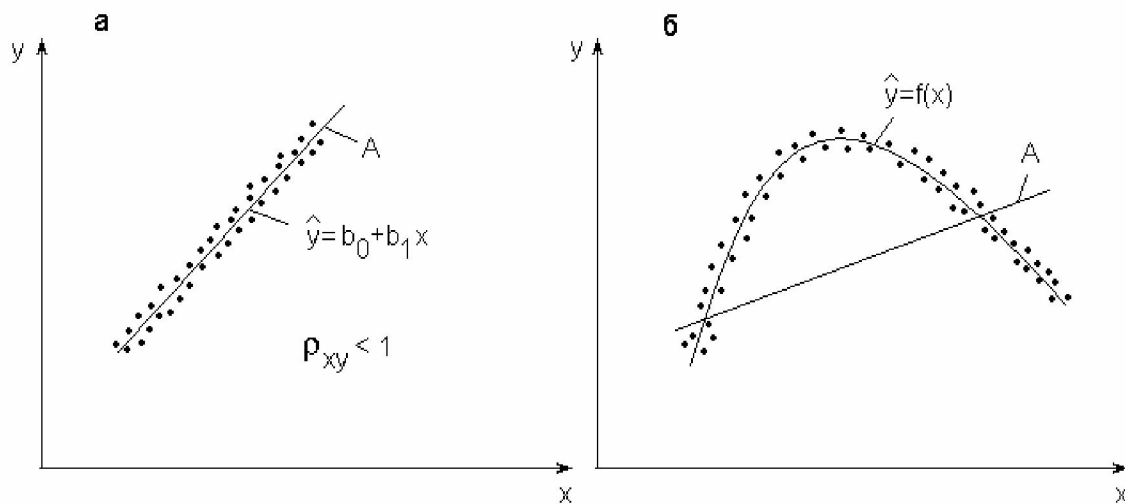


Рис.4.8. К понятию коэффициента парной корреляции

Даже при выполнении этих, вообще говоря достаточно жестких условий, не всякое значение выборочного коэффициента корреляции является достаточным для статистического обоснования выводов о наличии действительно надежной корреляционной связи между фактором и откликом. Надежность статистических характеристик ослабевает с уменьшением объема выборки (n). Так, при $n=2$ через две экспериментальные точки можно провести только одну прямую и зависимость будет функциональной, при этом выборочный коэффициент корреляции равен единице ($r_{xy}=1$). Однако это не означает надежность полученных статистических характеристик в силу весьма и весьма ограниченного объема выборки. Значит, вычислять коэффициент корреляции по результатам двух наблюдений бессмысленно, так как он заведомо будет равен единице, и это будет обусловлено не свойствами переменных и их взаимным отношением, а только числом наблюдений.

В связи с этим требуется проверка того, насколько значимо отличается выборочный коэффициент корреляции r_{xy} от его действительного значения r_{xy}^* . При достаточно большом объеме выборки $n \rightarrow \infty$ $r_{xy}^* = r_{xy}$.

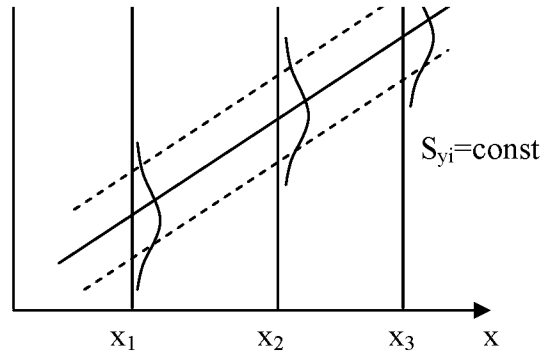


Рис.4.9. К понятию коэффициента парной корреляции в случае двумерного нормального распределения параметров

Таким образом, требуется проверка значимости выборочного коэффициента парной корреляции и оценка его доверительного интервала.

Для определения значимости r_{xy} сформулируем нуль-гипотезу $H_0: r_{xy}^* = 0$, т.е. корреляция отсутствует. Для этого рассчитывается экспериментальное значение t-критерия Стьюдента

$$t = |r_{xy}| \frac{\sqrt{n-2}}{\sqrt{1-(r_{xy})^2}} \quad (4.21)$$

и сравнивается с теоретическим при числе степеней свободы $n-2$.

Если $t \geq t_{\alpha; n-2}$ при заданном уровне значимости α , то нулевая гипотеза отклоняется, а альтернативная гипотеза $H_1: r_{xy}^* \neq 0$, о том, что коэффициент корреляции существенен, принимается.

Определение доверительного интервала коэффициента корреляции. При малых объемах выборки ($n < 20$) можно рекомендовать построение доверительного интервала для r_{xy}^* , которое основано на преобразовании Р.Фишера. Он предложил такое нелинейное преобразование величины r_{xy} , при котором закон распределения этой оценки, вообще говоря, довольно сложный, практически приближается к нормальному. Это преобразование производится по формуле

$$Z^* = \frac{1}{2} \ln \frac{1+r_{xy}}{1-r_{xy}}. \quad (4.22)$$

Среднеквадратичное отклонение случайной величины Z^* зависит от числа опытов

$$\sigma_{Z^*} = \frac{1}{\sqrt{n-3}}, \quad (4.23)$$

а математическое ожидание очень близко к числу, получающемуся после подстановки в формулу (4.22) вместо r_{xy} истинного значения коэффициента корреляции r_{xy}^* . Эти свойства величины Z^* позволяют просто оценить, в каких пределах может находиться истинное значение коэффициента корреляции, если по n опытам получены некоторые значения его выборочного значения (оценки) r_{xy} . Если граничное значение r_{xy} имеет тот же знак, что и r_{xy}^* , то можно считать в первом приближении, что корреляционная связь между переменными достоверна.

Пример 4.1. При обработке $n=17$ пар данных x и y выборочный коэффициент корреляции составил $r_{xy} = -0,94$, т.е. величина y связана с x достаточно сильной причинной связью, близкой к функциональной зависимости. Требуется определить значимость и найти доверительный интервал выборочного коэффициента корреляции.

Определение значимости коэффициента r_{xy}

$$t = |r_{xy}| \frac{\sqrt{n-2}}{\sqrt{1-(r_{xy})^2}} = 0,94 \frac{\sqrt{17-2}}{\sqrt{1-(0,94)^2}} = 10,6.$$

Критерий Стьюдента $t_{0,05;15} = 2,13$ (**СТЬЮДРАСПОБР** (0,05;15)=2,13145).

Поскольку $t > t_{\alpha;n-2}$, то коэффициент корреляции существенен.

Определение доверительного интервала. По формулам (4.22) и (4.23) определим величину Z^* :

$$Z^* = \frac{1}{2} \ln \frac{1-0,94}{1+0,94} = -1,738$$

и ее среднеквадратичное отклонение:

$$S_{Z^*} = \frac{1}{\sqrt{17-3}} = 0,267.$$

Зададимся вероятностью того, что истинное значение Z отличается от вычисленного на основании оценки коэффициента корреляции Z^* не более чем на δ_Z . Учитывая нормальный закон распределения Z , имеем при вероятности:

$$90\%: \quad \delta_Z = 1,64 \cdot S_Z = 1,67 \cdot 0,267 = 0,438;$$

$$95\%: \quad \delta_Z = 1,96 \cdot 0,267 = 0,523;$$

$$99,7\%: \quad \delta_Z = 3,00 \cdot 0,267 = 0,801.$$

Таким образом, истинное значение Z лежит в пределах $Z_1 \leq Z \leq Z_2$, где с вероятностью, например, 90%, $Z_1 = -1,738 - 0,438 = -2,176$ и $Z_2 = -1,738 + 0,438 = -1,300$. Для заданных значений вероятностей значения Z_1 и Z_2 составят:

$$90\%: \quad Z_1 = -2,176, \quad Z_2 = -1,300;$$

$$95\%: \quad Z_1 = -2,261, \quad Z_2 = -1,215;$$

$$99,7\%: \quad Z_1 = -2,539, \quad Z_2 = -0,937.$$

Этим значениям Z_1 и Z_2 соответствуют коэффициенты корреляции, полученные из формулы (4.22). Чтобы определить численные значения коэффициентов корреляции из формулы (4.22), можно воспользоваться инструментом «Подбор параметра» из электронных таблиц Microsoft Excel (меню «Сервис/Подбор параметра...»). В результате получим следующее решение:

$$90\%: \quad r_1 = -0,97, \quad r_2 = -0,86, \quad \text{т.е.} \quad -0,97 \leq r_{xy} \leq -0,86;$$

$$95\%: \quad r_1 = -0,98, \quad r_2 = -0,84, \quad \text{т.е.} \quad -0,98 \leq r_{xy} \leq -0,84;$$

$$99,7\%: \quad r_1 = -0,99, \quad r_2 = -0,73, \quad \text{т.е.} \quad -0,99 \leq r_{xy} \leq -0,73.$$

Следовательно, доверительные интервалы подтверждают достаточно сильную причинную связь между анализируемыми параметрами.

Таким образом, корреляционный анализ устанавливает связь между исследуемыми случайными переменными и оценивает тесноту этой связи.

4.5. Регрессионный анализ

Ниже излагаются основные положения регрессионного анализа, применение которого для обработки результатов наблюдений связано с меньшим числом ограничений, чем при корреляционном анализе. Как и корреляционный анализ, регрессионный анализ включает в себя построение уравнения регрессии, например, методом наименьших квадратов и статистическую оценку результатов. Если в регрессионном анализе расчет коэффициентов ведется теми же методами, например наименьших квадратов, то его теоретические предпосылки требуют других способов статистической оценки результатов.

При проведении регрессионного анализа примем следующие допущения:

1) входной параметр x измеряется с пренебрежимо малой ошибкой. Появление ошибки в определении y объясняется наличием в процессе не выяв-

ленных переменных и случайных воздействий, не вошедших в уравнение регрессии;

2) результаты наблюдений $y_1, y_2, \dots, y_i, \dots, y_n$ над выходной величиной представляют собой независимые нормально распределенные случайные величины;

3) при проведении эксперимента с объемом выборки n при условии, что каждый опыт повторен m^* раз, выборочные дисперсии $S_1^2, \dots, S_i^2, \dots, S_n^2$ должны быть однородны. При выполнении измерений в различных условиях возникает задача сравнения точности измерений. При этом следует подчеркнуть, что экспериментальные данные можно сравнивать только тогда, когда их дисперсии однородны. Это означает, как уже отмечалось (см. п. 3.5.1 и 3.5.2), принадлежность экспериментальных данных к одной и той же генеральной совокупности. Напомним: однородность дисперсий свидетельствует о том, что среди сравниваемых дисперсий нет таких, которые с заданной надежностью превышали бы все остальные, т.е. была бы большая ошибка. При одинаковом числе параллельных опытов однородность дисперсии, как мы уже показали, можно оценить по критерию Кохрена, а для сравнения двух дисперсий целесообразно воспользоваться F -критерием Фишера (см. примеры 3.4–3.5).

После того как уравнение регрессии найдено, необходимо провести статистический анализ результатов. Этот анализ состоит в следующем: проверяется значимость всех коэффициентов и устанавливается адекватность уравнения.

4.5.1. Проверка адекватности модели

При моделировании приходится формализовать связи исследуемого явления (процесса), из-за чего возможна потеря некоторой информации об объекте. Иногда некоторые связи не учитываются. В то же время основное требование к математической модели заключается в ее пригодности для решения поставленной задачи и адекватности процессу. Регрессионную модель называют адекватной, если предсказанные по ней значения y согласуются с результатами наблюдений. Так, построив модель в виде линейного уравнения регрессии, мы хотим, в частности, убедиться, что никакие другие модели не дадут значительного улучшения в описании предсказания значений y . В основе процедуры проверки адекватности модели лежат предположения, что случайные ошибки наблюдений являются независимыми, нормально распределенными случайными величинами с нулевыми средними значениями и одинаковыми дисперсиями.

Сформулируем нуль-гипотезу H_0 : "Уравнение регрессии адекватно".

Альтернативная гипотеза H_1 : "Уравнение регрессии неадекватно".

Для проверки этих гипотез принято использовать F-критерий Фишера.

При этом общую дисперсию (дисперсию выходного параметра) S_y^2 сравнивают с остаточной дисперсией $S_{y \text{ ост}}^2$.

Напомним, что

$$S_y^2 = \frac{\sum_{i=1}^n [y_i - \bar{y}]^2}{n-1};$$
$$S_{y \text{ ост}}^2 = \frac{\sum_{i=1}^n [y_i - \hat{y}_i]^2}{n-l},$$

(4.24)

где $l=k+1$ – число членов аппроксимирующего полинома, а k – число факторов. Так, например, для линейной зависимости (4.5) $k=1$, $l=2$.

В дальнейшем определяется экспериментальное значение F-критерия

$$F = S_y^2 / S_{y_{\text{ост}}}^2, \quad (4.25)$$

который в данном случае показывает, во сколько раз уравнение регрессии предсказывает результаты опытов лучше, чем среднее $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = C = \text{const}$.

Если $F > F_{\alpha; m_1; m_2}$, то уравнение регрессии адекватно. Чем больше значение F превышает $F_{\alpha; m_1; m_2}$ для выбранного α и числа степеней свободы $m_1 = n - 1$, $m_2 = n - l$, тем эффективнее уравнение регрессии.

Рассмотрим также случай, когда в каждой i -й точке x_i для повышения надежности и достоверности осуществляется не одно, а m^* параллельных измерений (примем для простоты, что m^* одинаково для каждого фактора). Тогда число экспериментальных значений величины y составит $n_{\Sigma} = n \cdot m^*$.

В этом случае оценка адекватности модели производится следующим образом:

1) определяется $\bar{y}_i = \sum_{j=1}^{m^*} y_{ij} / m^*$ – среднее из серии параллельных опытов

при $x = x_i$, где y_{ij} – значение параметра y при $x = x_i$ в j -м случае;

2) рассчитываются значения параметра \hat{y}_i по уравнению регрессии при $x = x_i$;

3) рассчитывается дисперсия адекватности

$$S_{\text{ад}}^2 = \frac{m^* \sum_{i=1}^n [\bar{y}_i - \hat{y}_i]^2}{n - l},$$

где n – число значений x_i ; l – число членов аппроксимирующего полинома (коэффициентов b_i), для линейной зависимости $l = 2$;

4) определяется выборочная дисперсия Y при $x = x_i$:

$$S_i^2 = \frac{\sum_{j=1}^{m^*} [y_{ij} - \bar{y}_i]^2}{m^* - 1};$$

5) определяется дисперсия воспроизводимости

$$S_{\text{восп}}^2 = \sum_{i=1}^n S_i^2 / n.$$

Число степеней свободы этой дисперсии равно $m=n(m^*-1)$;

6) определяется экспериментальное значение критерия Фишера

$$F = S_{\text{ад}}^2 / S_{\text{восп}}^2.$$

7) определяется теоретическое значение этого же критерия $F_{\alpha; m_1; m_2}$,

где $m_1=n-l$; $m_2= n (m^*-1)$;

8) если $F \leq F_{\alpha; m_1; m_2}$, то уравнение регрессии адекватно, в противном случае – нет.

4.5.2. Проверка значимости коэффициентов уравнения регрессии

Надежность оценок b_i уравнения регрессии можно охарактеризовать их доверительными интервалами Δb_i , в которых с заданной вероятностью находится истинное значение этого параметра.

Наиболее просто построить доверительные интервалы для параметров линейного уравнения регрессии, т.е. коэффициентов b_0 и b_1 . При этом предполагается, что для каждого значения случайной величины $x=x_i$ имеется распределение со средним значением $\hat{y}_i = b_0 + b_1 x_i$ и дисперсией $S_{y_i}^2 = S_{\text{восп}}^2$. Иными словами, делается допущение, что случайная величина Y распределена нормально при каждом значении x_i , а дисперсия $S_{y_i}^2$ во всем интервале изменения x постоянна: $S_{y_i}^2 = \text{const}$ (см. рис. 4.9).

Для линейного уравнения среднеквадратичное отклонение i -го коэффициента уравнения регрессии S_{b_i} можно определить по закону накопления ошибок

$$S_{b_i} = \sqrt{\sum_{j=1}^n \left(\frac{\partial b_j}{\partial y_i} \right)^2 S_j^2}. \quad (4.26)$$

При условии, что $S_{y1}^2 = S_{y2}^2 = \dots = S_{yi}^2 = \dots = S_{yn}^2 = S_{\text{восп}}^2$, получим

$$S_{b_0} = \sqrt{\frac{S_{\text{восп}}^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}}; \quad (4.27)$$

$$S_{b_1} = \sqrt{\frac{S_{\text{восп}}^2 \cdot n}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}}. \quad (4.27a)$$

S_{b_0} и S_{b_1} называются соответственно стандартной ошибкой свободного члена и стандартной ошибкой коэффициента регрессии.

Проверка значимости коэффициентов выполняется по критерию Стьюдента. При этом проверяется нуль-гипотеза $H_0: b_i=0$, т.е. i -й коэффициент генеральной совокупности при заданном уровне значимости α отличен от нуля.

Построим доверительный интервал для коэффициентов уравнения регрессии

$$\Delta b_i = t_{\alpha; n-l} \cdot S_{b_i}, \quad (4.28)$$

где число степеней свободы в критерии Стьюдента определяется по соотношению $n-l$. Потеря $l=k+1$ степеней свободы обусловлена тем, что все коэффициенты b_i рассчитываются зависимо друг от друга, что следует из уравнений (4.16) и (4.16a).

Тогда доверительный интервал для Δb_i коэффициента уравнения регрессии составит $(b_i - \Delta b_i; b_i + \Delta b_i)$. Чем уже доверительный интервал, тем с большей уверенностью можно говорить о значимости этого коэффициента.

Необходимо всегда помнить рабочее правило: "Если абсолютная величина коэффициента регрессии больше, чем его доверительный интервал, то этот коэффициент значим".

Таким образом, если $|b_i| > |\Delta b_i|$, то b_i коэффициент значим, в противном случае – нет.

Незначимые коэффициенты исключаются из уравнения регрессии, а оставшиеся коэффициенты пересчитываются заново, так как они зависимы и в формулы для их расчета (4.16) и (4.16а) входят разноименные переменные.

4.6. Линейная множественная регрессия

При изучении множественной регрессии не существует графической интерпретации многофакторного пространства. При проведении экспериментов в такой ситуации исследователь записывает показания приборов о состоянии функции отклика y и всех факторов x_i , от которых она зависит. Результат исследований – это матрица наблюдений.

$$\begin{pmatrix} y_1 & x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1k} \\ y_2 & x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ y_i & x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ik} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ y_n & x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{nk} \end{pmatrix}. \quad (4.29)$$

Здесь n – число опытов; k – число факторов; x_{ij} – значение j -го фактора в i -м опыте; y_i – значение выходного параметра для i -го опыта.

Задача линейной множественной регрессии состоит в построении гиперплоскости в $(k+1)$ -мерном пространстве, отклонения результатов наблюдений y_i от которой были бы минимальными при использовании метода наименьших квадратов. Или, другими словами, следует определить значения коэффициентов $b_0, \dots, b_j, \dots, b_k$ в линейном полиноме

$$\hat{y} = b_0 + \sum_{j=1}^k b_j x_j,$$

минимизирующие выражение

$$\Phi = \sum_i^n (y_i - \hat{y}_i)^2 = \sum_i^n [y_i - (b_0 + b_1x_1 + \dots + b_jx_{ij} + \dots + b_kx_{nk})]^2 \rightarrow \min_{b_j}. \quad (4.30)$$

Процедура определения коэффициентов $b_0, \dots, b_j, \dots, b_k$ в принципе не отличается от одномерного случая, рассмотренного ранее, и поэтому здесь не приводится.

Для оценки тесноты связи между функцией отклика \hat{y} и несколькими факторами $x_1, x_2, \dots, x_j, \dots, x_k$ используют коэффициент множественной корреляции R , который всегда положителен и изменяется в пределах от 0 до 1. Чем больше R , тем качественнее предсказания данной моделью опытных данных с точки зрения близости ее к функциональной. При функциональной линейной зависимости $R=1$.

Расчеты обычно начинают с вычисления парных коэффициентов корреляции, при этом вычисляются два типа парных коэффициентов корреляции:

- 1) r_{yx_j} – коэффициенты, определяющие тесноту связи между функцией отклика \hat{y} и одним из факторов x_j ;
- 2) $r_{x_jx_u}$ – коэффициенты, показывающие тесноту связи между одним из факторов x_j и фактором x_u ($j, u = 1 \div k$).

Если один из коэффициентов $r_{x_jx_u}$ окажется равным 1, то это означает, что факторы x_j и x_u функционально связаны между собой. Тогда целесообразно один из них исключить из рассмотрения, причем оставляют тот фактор, у которого коэффициент r_{yx_j} больше.

После вычисления всех парных коэффициентов корреляции можно построить матрицу коэффициентов корреляции следующего вида:

$$\begin{pmatrix}
1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_j} & \dots & r_{yx_k} \\
r_{x_1y} & 1 & r_{x_1x_2} & \dots & r_{x_1x_j} & \dots & r_{x_1x_k} \\
r_{x_2y} & r_{x_2x_1} & 1 & \dots & r_{x_2x_j} & \dots & r_{x_2x_k} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
r_{x_jy} & r_{x_jx_1} & r_{x_jx_2} & \dots & 1 & \dots & r_{x_jx_k} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
r_{x_ky} & r_{x_kx_1} & r_{x_kx_2} & \dots & r_{x_kx_j} & \dots & 1
\end{pmatrix} \quad (4.31)$$

Однако парные коэффициенты корреляции не характеризуют тесноту связи, так как они вычисляются при случайно изменяющихся значениях других факторов. Действительно, при рассмотрении трех и более случайных величин коэффициенты корреляции любой пары из этих случайных величин могут не дать правильного представления о степени связи между всеми случайными величинами. Это объясняется тем, что на закон распределения вероятностей исследуемой пары случайных величин могут оказывать влияние и другие рассматриваемые случайные величины. Это обстоятельство делает необходимым введение показателей стохастической связи между парой случайных величин при условии, что значения других случайных величин зафиксированы. В этом случае говорят о статистическом анализе частных связей. Используя матрицу (4.31), можно вычислить частные коэффициенты корреляции, которые показывают степень влияния одного из факторов x_j на функцию отклика \hat{y} при условии, что остальные факторы остаются на постоянном уровне. Формула для вычисления частных коэффициентов корреляции имеет вид

$$r_{yx_1, x_2, \dots, x_j, \dots, x_k} = D_{1j} / \sqrt{D_{11} \cdot D_{jj}}, \quad (4.32)$$

где D_{1j} – определитель матрицы, образованной из матрицы (4.31) вычеркиванием 1-й строки и j -го столбца. Определители D_{11} и D_{jj} вычисляются аналогично. Как и парные коэффициенты, частные коэффициенты корреляции изменяются от -1 до +1.

Значимость и доверительный интервал для коэффициентов частной корреляции определяются так же, как для коэффициентов парной корреляции, только число степеней свободы вычисляют по формуле

$$m = n - k - 1, \quad (4.33)$$

где $k = k - 1$ – порядок частного коэффициента парной корреляции.

Для вычисления коэффициента множественной корреляции $R_{y, x_1, x_2, \dots, x_j, \dots, x_k}$ используют матрицу (4.31):

$$R_{y, x_1, x_2, \dots, x_j, \dots, x_k} = \sqrt{1 - D/D_{11}}, \quad (4.34)$$

где D – определитель матрицы (4.31).

Множественный коэффициент корреляции дает оценку тесноты связи между y и совокупностью всех переменных $x_1, x_2, \dots, x_j, \dots, x_k$.

Если число опытов n сравнимо с числом коэффициентов $l = k + 1$, связи оказываются преувеличенными. Поэтому следует исключить систематическую погрешность, физический смысл которой состоит в следующем. Если разность n и l будет уменьшаться, то коэффициент множественной корреляции R будет возрастать и при $n - l = 0$ окажется равным $R = +1$, а уравнение регрессии превратится в функциональное уравнение гиперплоскости, которая пройдет через все n экспериментальных точек. Однако ясно, что случайный характер переменных процесса при этом не может измениться. В связи с этим требуется оценка значимости коэффициента множественной корреляции.

Значимость коэффициента множественной корреляции проверяется по критерию Стьюдента:

$$t = \frac{R}{\bar{S}_R} \geq t_{\alpha, m}; \quad m = n - k - 1,$$

где \bar{S}_R – среднеквадратичная погрешность коэффициента множественной корреляции, рассчитываемая по выражению

$$\bar{S}_R = (1 - R^2) \sqrt{n - k - 1}. \quad (4.35)$$

Значимость R можно проверить также по критерию Фишера

$$F = \frac{R^2(n - k - 1)}{(1 - R^2)k}. \quad (4.36)$$

Если расчетное значение F превышает теоретическое $F_{\alpha; m_1; m_2}$, то гипотезу о равенстве коэффициента множественной корреляции нулю отвергают и связь считают статистически значимой. Теоретическое (табличное) значение критерия Фишера определяется для выбранного уровня значимости α и числа степеней свободы $m_1 = n - k - 1$ и $m_2 = k$.

Если коэффициент множественной корреляции оказался неожиданно малым, хотя априорно известно, что между выходом y и входами x_1, \dots, x_k должна существовать достаточно тесная корреляционная связь, то возможными причинами такого явления могут быть следующие:

а) ряд существенных факторов не учтен, и следует включить в рассмотрение дополнительно эти существенные входные параметры;

б) линейное уравнение плохо аппроксимирует в действительности нелинейную зависимость $\hat{y} = f(x_1, \dots, x_k)$, и следует определить коэффициенты уже нелинейного уравнения регрессии методами регрессионного анализа;

в) рабочий диапазон рассматриваемых факторов находится в районе экстремума функции отклика – в этом случае следует расширить диапазон изменения входных переменных, а также перейти к нелинейной математической модели объекта.

4.7. Нелинейная регрессия

Используя подходы, изложенные ранее, можно построить практически любые формы нелинейной связи. С этой целью в инженерной практике очень часто используют линеаризующие преобразования.

В табл. 4.1 приведены часто встречающиеся парные зависимости и линеаризующие преобразования переменных. Качество преобразования результатов проверяют с помощью уравнения $\hat{y} = b_0' + b_1' \cdot x'$.

Таблица 4.1

Функции и линеаризующие преобразования

№ п/п	Функция	Линеаризующие преобразования			
		Преобразование переменных		Выражения для величин b_0 и b_1	
		y'	x'	b_0'	b_1'
1	$y = b_0 + b_1/x$	y	$1/x$	b_0	b_1
2	$y = 1/(b_0 + b_1x)$	$1/y$	x	b_0	b_1
3	$y = x/(b_0 + b_1x)$	x/y	x	b_0	b_1
4	$y = b_0 b_1^x$	$\lg(y)$	x	$\lg(b_0)$	$\lg(b_1)$
5	$y = b_0 \cdot e^{b_1x}$	$\ln(y)$	x	$\ln(b_0)$	b_1
6	$y = 1/(b_0 + b_1e^{-x})$	$1/y$	e^{-x}	b_0	b_1
7	$y = b_0x^{b_1}$	$\lg(y)$	$\lg(x)$	$\lg(b_0)$	b_1
8	$y = b_0 + b_1 \lg(x)$	y	$\lg(x)$	b_0	b_1
9	$y = b_0/(b_1 + x)$	$1/y$	x	b_1/b_0	$1/b_0$
10	$y = b_0x/(b_1 + x)$	$1/y$	$1/x$	b_1/b_0	$1/b_0$
11	$y = b_0e^{b_1/x}$	$\ln(y)$	$1/x$	$\ln(b_0)$	b_1
12	$y = b_0 + b_1x^n$	y	x^n	b_0	b_1

После вычисления коэффициентов b_0' и b_1' , так же как в случае линейной зависимости от одного фактора, выполняют обратные преобразования, т.е. по b_0' и b_1' определяют b_0 и b_1 . Аналогичный подход обычно используют и при множественном нелинейном регрессионном анализе.

5. ОЦЕНКА ПОГРЕШНОСТЕЙ РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ

5.1. Оценка погрешностей определения величин функций

При изложении материалов, касающихся оценки погрешностей результатов наблюдений, будем в дальнейшем придерживаться методологии решения этих задач, представленной в учебном пособии [2].

Необходимость в определении погрешности величин функций по известным значениям погрешностей их аргументов (факторов) возникает при оценке точности результатов математического эксперимента, а также результатов так называемых косвенных измерений. Под косвенным измерением понимается такое, в результате которого значение искомой величины y рассчитывают по известной зависимости ее от других величин x_1, x_2, \dots, x_k , измеренных другим способом, т.е.

$$\hat{y} = f(x_1, x_2, \dots, x_i, \dots, x_k), \quad (5.1)$$

где $x_1, x_2, \dots, x_i, \dots, x_k$ – аргументы, определенные независимо друг от друга. В дальнейшем будем полагать, что погрешности определения величины y обусловлены лишь неточностью численных значений величин $x_1, x_2, \dots, x_i, \dots, x_k$, входящих под знак функции.

Обозначим истинное значение i -го параметра через x_i , среднее значение – через \bar{x}_i , а абсолютную погрешность его измерения – через Δx_i . Разложим функцию $f(x_1, x_2, \dots, x_k)$ в ряд Тейлора, сохраняя члены с нулевой и первыми степенями погрешностей:

$$f(x_1, \dots, x_i, \dots, x_k) = f(\bar{x}_1, \dots, \bar{x}_i, \dots, \bar{x}_k) + \sum_{i=1}^k \left(\frac{\partial f(x_1, \dots, x_i, \dots, x_k)}{\partial x_i} \right)_{x_i = \bar{x}_i} \cdot \Delta x_i,$$

где все производные $\left(\frac{\partial f(x_1, \dots, x_i, \dots, x_k)}{\partial x_i} \right)_{x_i = \bar{x}_i}$ вычислены при значениях $x_i = \bar{x}_i$.

Тогда

$$\Delta y^2 \approx \left[\overline{f(x_1, \dots, x_i, \dots, x_k) - f(\bar{x}_1, \dots, \bar{x}_i, \dots, \bar{x}_k)} \right]^2 = \left[\sum_{i=1}^k \left(\frac{\partial f(x_1, \dots, x_i, \dots, x_k)}{\partial x_i} \right)_{x_i = \bar{x}_i} \Delta x_i \right]^2 \approx$$

$$\approx \sum_{i=1}^k \left(\frac{\partial f(x_1, \dots, x_i, \dots, x_k)}{\partial x_i} \right)_{x_i = \bar{x}_i}^2 \Delta x_i^2 = \sum_{i=1}^k \Delta y_i^2,$$

где $\Delta y_i = \left(\frac{\partial f(x_1, \dots, x_k)}{\partial x_i} \right)_{x_i = \bar{x}_i} \cdot \Delta x_i.$ (5.2)

Следовательно, Δy_i – это составляющие погрешности функции, обусловленные погрешностью i -го аргумента x_i .

Доверительная вероятность, соответствующая величине Δy_i , численно равна доверительной вероятности, с которой найдена погрешность Δx_i .

Для относительной погрешности вместо соотношения (5.2) используют выражение

$$\Delta_{y_i}^* = \frac{1}{f} \frac{\partial f}{\partial x_i} \Delta x_i = \frac{\partial \ln(f)}{\partial x_i} \Delta x_i. \quad (5.3)$$

Соотношения (5.2) и (5.3) применимы для расчета как случайных, так и систематических погрешностей.

Общая абсолютная (Δy_Σ) и относительная (Δ_Σ^*) погрешности определения функции могут быть найдены с помощью выражений

$$\Delta y_\Sigma = \pm \sqrt{\sum_{i=1}^k (\Delta y_i)^2}; \quad (5.4)$$

$$\Delta_\Sigma^* = \pm \sqrt{\sum_{i=1}^k (\Delta_{y_i}^*)^2}. \quad (5.5)$$

Предполагается, что все составляющие имеют нормальный закон распределения.

Частные производные, входящие в соотношения (5.2) и (5.3), не всегда могут быть найдены аналитически. Часто не удается разрешить искомую задачу относительно искомой величины y в явном виде. В этих случаях полезно использовать численные методы определения производных.

Пример 5.1. Рассмотрим погрешность определения массового расхода газового потока стандартным сужающим устройством. При этом будем считать, что случайная составляющая погрешности отсутствует, а поправка на сжимаемость потока равна единице.

Тогда с учетом выражения для определения массового расхода вещества

$$G = \alpha F_0 \varepsilon \sqrt{2\rho(p_1 - p_2)} = \alpha F_0 \varepsilon \sqrt{2\rho h}, \quad (5.6)$$

где F_0 – площадь сужающего устройства; ε – поправочный множитель на сжимаемость вещества, расход которого измеряется ($\varepsilon = 1$); ρ – плотность потока перед сужающим устройством; h – перепад статического давления на сужающем устройстве, α – коэффициент расхода.

Используя соотношения (5.2) и (5.4), получим следующие формулы для расчета абсолютной и относительной погрешности определения расхода:

$$\Delta G = \pm \sqrt{\left(\frac{\partial G}{\partial \rho} \Delta \rho\right)^2 + \left(\frac{\partial G}{\partial h} \Delta h\right)^2}; \quad (5.7)$$

$$\Delta_G^* = \pm \sqrt{\left(\frac{\partial G}{\partial \rho} \frac{\Delta \rho}{G}\right)^2 + \left(\frac{\partial G}{\partial h} \frac{\Delta h}{G}\right)^2} = \pm \frac{1}{2} \sqrt{\left(\frac{\Delta \rho}{\rho}\right)^2 + \left(\frac{\Delta h}{h}\right)^2}, \quad (5.8)$$

где $\frac{\partial G}{\partial \rho} = \alpha F_0 \varepsilon \sqrt{\frac{h}{2\rho}}$; $\frac{\partial G}{\partial h} = \alpha F_0 \varepsilon \sqrt{\frac{\rho}{2h}}$.

Учтем далее погрешности определения плотности потоков. В соответствии с уравнением состояния газа $\rho = p/RT$, где p и T – соответственно абсолютное давление и температура газа перед сужающим устройством, R – универсальная газовая постоянная. Абсолютная погрешность определения плотности потока без учета погрешности газовой постоянной составит

$$\Delta \rho = \pm \sqrt{\left(\frac{\partial \rho}{\partial p} \Delta p\right)^2 + \left(\frac{\partial \rho}{\partial T} \Delta T\right)^2}, \quad (5.9)$$

где $\frac{\partial \rho}{\partial p} = \frac{1}{RT}$; $\frac{\partial \rho}{\partial T} = -\frac{p}{RT^2}$;

относительная погрешность

$$\Delta_{\rho}^* = \frac{\Delta\rho}{\rho} = \pm \sqrt{\left(\frac{\Delta p}{p}\right)^2 + \left(\frac{\Delta T}{T}\right)^2}. \quad (5.10)$$

Тогда относительная погрешность определения массового расхода газового потока будет

$$\Delta_G^* = \pm \frac{1}{2} \sqrt{\left(\frac{\Delta p}{p}\right)^2 + \left(\frac{\Delta T}{T}\right)^2 + \left(\frac{\Delta h}{h}\right)^2}. \quad (5.11)$$

Здесь p , T , h – значения измеренных параметров; Δp , ΔT , Δh – их абсолютные погрешности. Численные значения Δp , ΔT , Δh определяются в основном инструментальной погрешностью и могут быть вычислены с учетом класса точности используемых приборов для измерения p и h . Погрешность измерения T определяется с учетом вида измерительного устройства температуры.

Абсолютная погрешность определения массового расхода газового потока

$$\Delta G = G \Delta_G^*, \quad (5.12)$$

где G – значение расхода, измеренное экспериментально. Таким образом, истинное значение расхода будет равно

$$G_{ист} = G \pm \Delta G, \quad (5.13)$$

5.2. Обратная задача теории экспериментальных погрешностей

Целью обратной задачи является определение погрешностей величин-аргументов, если известны погрешности функций и вид функциональной зависимости. Необходимость в решении таких задач возникает при выборе того или иного комплекса измерительной аппаратуры или метода определения искомой величины, позволяющих найти значение этой величины с определенной погрешностью.

Обратная задача в общем случае является неопределенной, поскольку имеется одно уравнение с k неизвестными. Иначе говоря, удовлетворить условию задачи можно при различных комбинациях значений погрешностей аргументов. [2].

Очень часто удовлетворительное решение обратной задачи оказывается возможным при использовании так называемого принципа равных влияний. Он заключается в том, что при решении задачи накладывается дополнительное требование, чтобы все члены в правой части выражений (5.4) и (5.5) оказывали одинаковое влияние на погрешности функции.

Применяя принцип равных влияний к относительной погрешности функции, определяемой соотношением (5.5), получим [2]

$$(\Delta_{y_1}^*)^2 = (\Delta_{y_2}^*)^2 = \dots = (\Delta_{y_k}^*)^2 = \frac{(\Delta_\Sigma^*)^2}{k}, \quad (5.14)$$

$$\Delta_{y_i}^* = \frac{\Delta_\Sigma^*}{\sqrt{k}}. \quad (5.15)$$

С учетом (5.3) легко получить выражение для определения абсолютных Δx_i и относительных Δx_i^* погрешностей всех аргументов:

$$\Delta x_i = \frac{\Delta_\Sigma}{\sqrt{k}} \frac{1}{\frac{\partial \ln [f(x_1, x_2, \dots, x_k)]}{\partial x_i}}; \quad (5.16)$$

$$\Delta_{x_i}^* = \frac{\Delta_\Sigma}{x_i \sqrt{k}} \frac{1}{\frac{\partial \ln [f(x_1, x_2, \dots, x_k)]}{\partial x_i}}. \quad (5.17)$$

В дальнейшем могут иметь место три возможных случая:

- значения погрешностей всех аргументов таковы, что лежат в пределах точности, доступной при измерениях с помощью имеющихся средств измерений;
- значения некоторых погрешностей настолько малы, что обеспечить соответствующую точность с помощью имеющихся средств измерений не представляется возможным;

- значения всех погрешностей малы, и обеспечить такую точность невозможно.

В первом случае проблем не возникает и поставленная задача имеет решение. Во втором случае прежде всего следует попытаться решить задачу путем увеличения погрешности тех аргументов, у которых оказалось невозможным обеспечить требуемую первоначальную точность измерений при одновременном уменьшении погрешностей остальных аргументов.

Если этот путь не дает приемлемых результатов, то остается один выход, связанный с поиском другого метода определения величины x . Этот выход является единственно возможным и для случая, когда значения погрешностей всех аргументов настолько малы, что обеспечить требуемую их точность с помощью имеющихся средств измерений не представляется возможным. При выборе другого метода измерений меняется вид функции $\hat{y} = f(\mathbf{X})$, а следовательно, меняются аргументы и значения их погрешностей.

Пример 5.2. Пусть требуется определить объем цилиндра диаметром $d=20$ мм и высотой $h=50$ мм с относительной погрешностью $\Delta_V^* = 0,01$. Найдем погрешности измерения величин d и h , соответствующие этому же значению доверительной вероятности, при которых исходная задача будет разрешена.

Учитывая, что объем цилиндра $V = \frac{\pi d^2 h}{4}$ и приняв закон распределения нормальным, с помощью соотношения (5.16) найдем

$$\Delta d = \pm \frac{\Delta_V^*}{\sqrt{2}} \frac{1}{\frac{\partial \ln(\pi d^2 h / 4)}{\partial d}} = \pm \frac{\Delta_V^*}{\sqrt{2}} \frac{d}{2} = \pm 0,07 \text{ мм};$$

$$\Delta h = \pm \frac{\Delta_V^*}{\sqrt{2}} \frac{1}{\frac{\partial \ln(\pi d^2 h / 4)}{\partial h}} = \pm \frac{\Delta_V^*}{\sqrt{2}} h = \pm 0,035 \text{ мм}.$$

5.3. Определение наивыгоднейших условий эксперимента

Под наивыгоднейшими условиями эксперимента понимаются такие, для которых погрешность результата эксперимента при фиксированном значении доверительной вероятности имеет наименьшее значение.

6. МЕТОДЫ ПЛАНИРОВАНИЯ ЭКСПЕРИМЕНТОВ. ЛОГИЧЕСКИЕ ОСНОВЫ

6.1. Основные определения и понятия

Ранее мы рассматривали пассивный эксперимент, и математическая статистика использовалась, в частности, при обработке экспериментальных данных. На стадии постановки эксперимента она не применялась. При активном же эксперименте математическая статистика используется уже на стадии постановки и планирования эксперимента.

Пассивный эксперимент предусматривает накопление информации "в режиме нормальной эксплуатации", но это требует много времени и затрат. Поэтому предлагается "не ждать милостей от природы", а активно вмешиваться в ход технологического процесса: «разбалтывать» («покачивать») его тихонько, но целенаправленно, и быстро накапливать при этом информацию. Программа покачивания как раз и задается планом. Сам метод планирования может изменяться в зависимости от вида задачи, но принцип покачивания остается.

Теория планирования эксперимента началась с работ знаменитого английского ученого Р.Фишера в 30-х годах XX столетия, использовавшего ее для решения агробиологических задач. В дальнейшем это направление было развито в пятидесятых годах в США Дж.Боксом и его сотрудниками. Отечественные ученые также внесли большой вклад в развитие теории эксперимента, предложив ряд новых методов, а инженеры-исследователи все шире применяют эти методы на практике.

Под математической теорией планирования эксперимента будем понимать науку о способах составления экономичных экспериментальных планов, которые позволяют извлекать наибольшее количество информации об объекте исследования, о способах проведения эксперимента, о способах обработки экспериментальных данных и их использования для оптимизации производственных процессов, а также инженерных расчетов.

Принятая терминология — это либо перевод терминов с английского, либо просто их перенос в оригинале, и это необходимо иметь в виду при чтении литературы по теории планирования экспериментов.

Истинный вид функции отклика $y=f(x_1, \dots, x_i, \dots, x_k)$ до эксперимента чаще всего неизвестен, в связи с чем для математического описания поверхности отклика используют уравнение

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i,u=1}^k \beta_{iu} x_i x_u + \sum_{i=1}^k \beta_{ii} x_i^2 + \dots, \quad (6.1)$$

где x_i, x_u – переменные факторы при $i=1, \dots, k; u=1, \dots, k; i \neq u$;

$$\beta_i = \left(\frac{\partial f}{\partial x_i} \right)_0; \beta_{iu} = \left(\frac{\partial^2 f}{\partial x_i \partial x_u} \right)_0; \beta_{ii} = \left(\frac{\partial^2 f}{2 \partial x_i^2} \right)_0 \text{ – коэффициенты.}$$

Это уравнение является разложением в ряд Тейлора неизвестной функции отклика в окрестности точки с $x_i=x_{i0}$.

На практике по результатам эксперимента производится обработка данных по методу наименьших квадратов. Этот метод позволяет найти оценку b коэффициентов β , и данный полином заменяется уравнением вида

$$\hat{y} = b_0 + \sum_{i=1}^k b_i x_i + \sum_{i,u=1}^k b_{iu} x_i x_u + \sum_{i=1}^k b_{ii} x_i^2 + \dots, \quad (6.2)$$

которое является регрессионной моделью (моделью регрессионного анализа). В этом выражении \hat{y} означает модельное, т.е. рассчитываемое по уравнению модели, значение выхода. Коэффициенты регрессии определяются экспериментально и служат для статистической оценки теоретических коэффициентов, т.е.

$$b_0 \rightarrow \beta_0, b_i \rightarrow \beta_i, b_{iu} \rightarrow \beta_{iu}, b_{ii} \rightarrow \beta_{ii}.$$

В регрессионной модели члены второй степени $x_i x_u, x_i^2$ характеризуют кривизну поверхности отклика. Чем больше кривизна этой поверхности, тем больше в модели регрессии членов высшей степени. На практике чаще всего стремятся ограничиться линейной моделью.

Последовательность активного эксперимента заключается в следующем:

1) разрабатывается схема проведения исследований, т.е. выполняется планирование эксперимента. При планировании экспериментов обычно требуется с наименьшими затратами и с необходимой точностью либо построить регрессионную модель процесса, либо определить его оптимальные условия;

2) осуществляется реализация опыта по заранее составленному исследователем плану, т.е. осуществляется сам активный эксперимент;

3) выполняется обработка результатов измерений, их анализ и принятие решений.

Таким образом, планирование эксперимента – это процедура выбора условий проведения опытов, их количества, необходимых и достаточных для решения задач с поставленной точностью.

Использование теории планирования эксперимента обеспечивает:

1) минимизацию, т.е. предельное сокращение необходимого числа опытов;

2) одновременное варьирование всех факторов;

3) выбор четкой стратегии, что позволяет принимать обоснованные решения после каждой серии опытов;

4) минимизацию ошибок эксперимента за счет использования специальных проверок.

Для иллюстрации некоторых из этих положений воспользуемся ставшим уже классическим примером из книги В.В.Налимова, Т.И.Голиковой [7].

6.2. Пример хорошего и плохого эксперимента

Рассмотрим пример – взвешивание трех объектов А, В, С на аналитических весах. Первый – традиционный – подход предусматривает последовательное взвешивание каждого из образцов. Исследователь вначале делает холостое взвешивание для определения нулевой точки весов, а затем по очереди взвешивает каждый из образцов. Это пример традиционного использования однофакторного эксперимента, т.е. здесь исследователь изучает реакцию на

поведение каждого из факторов в отдельности. Традиционная схема взвешивания трех объектов представлена в табл. 6.1.

Масса каждого объекта оценивается только по результатам двух опытов: того опыта, в котором на весы был положен изучаемый объект, и холостого опыта. Например, масса объекта А: $m_A = y_1 - y_0$. Как обычно, ошибка взвешивания предполагается независимой от взвешиваемой величины, аддитивной и имеющей одно и то же распределение. Тогда дисперсия измерения веса образца следующая:

$$\sigma_A^2 = \sigma_{y_1}^2 + \sigma_{y_0}^2 = 2\sigma^2, \quad (6.3)$$

где σ^2 — дисперсия любого взвешивания. Такими же будут и дисперсии весов образцов В и С.

Таблица 6.1

*Традиционное проведение эксперимента**

Номер опыта	А	В	С	Результат взвешивания
1	-1	-1	-1	y_0
2	+1	-1	-1	y_1
3	-1	+1	-1	y_2
4	-1	-1	+1	y_3

*) Когда образец кладется на весы, в таблице ставится +1, когда он на весах отсутствует, то -1.

Приведем теперь тот же эксперимент по несколько иной схеме, задаваемой матрицей планирования, приведенной в табл.6.2.

Таблица 6.2

Планирование эксперимента при взвешивании трех объектов

Номер опыта	А	В	С	Результат взвешивания
1	+1	-1	-1	y_1
2	-1	+1	-1	y_2
3	-1	-1	+1	y_3
4	+1	+1	+1	y_4

В первых трех опытах последовательно взвешивают объекты А, В, С, в последнем опыте тоже взвешивают объекты А, В, С, но все три объекта вместе, а "холостое" взвешивание не производится.

Легко заметить, что масса каждого объекта будет задаваться формулами

$$\begin{aligned}m_A &= \frac{1}{2}(y_1 - y_2 - y_3 + y_4); \\m_B &= \frac{1}{2}(y_2 - y_1 - y_3 + y_4); \\m_C &= \frac{1}{2}(y_3 - y_1 - y_2 + y_4).\end{aligned}\tag{6.4}$$

Масса объекта А, вычисленная по приведенной выше формуле, оказывается не искаженной массами весов объектов В и С, так как масса каждого из них входит в формулу для массы А дважды с разными знаками.

Найдем теперь дисперсию, связанную с ошибкой взвешивания, по новой схеме постановки экспериментов:

$$\sigma_A^2 = \frac{1}{4}(\sigma_{y_1}^2 + \sigma_{y_2}^2 + \sigma_{y_3}^2 + \sigma_{y_4}^2) = \sigma^2.\tag{6.5}$$

Аналогичным образом находим:

$$\sigma_B^2 = \sigma^2, \quad \sigma_C^2 = \sigma^2.$$

Мы видим, что при новой схеме дисперсия взвешивания получается вдвое меньше, чем при традиционной схеме, хотя в обоих случаях на взвешивание трех объектов затрачивалось по четыре опыта.

Зададимся вопросом: "В результате чего происходит увеличение точности экспериментов в два раза?".

В первом случае эксперимент был поставлен так, что каждую массу мы получали лишь из двух взвешиваний. При новой схеме взвешивания каждая масса вычислялась уже по результатам всех четырех взвешиваний. Вторую схему можно назвать многофакторной, поскольку здесь оперируют всеми факторами так, что каждая масса вычислялась по результатам сразу всех опытов, проведенных в данной серии экспериментов, – вот главная причина уменьшения дисперсии вдвое.

Не подумайте, что мы зря потратили время на обсуждение такой тривиальной задачи. Точно такой же подход используется при изучении других, более сложных задач.

Таким образом, использование теории планирования эксперимента может явиться одним из путей существенного повышения эффективности многофакторных экспериментальных исследований.

В планировании экспериментов применяются в основном планы первого и второго порядков. Планы более высоких порядков используются в инженерной практике редко. В связи с этим далее приводится краткое изложение методики составления планов эксперимента для моделей первого и второго порядков.

Под планами первого порядка понимают такие планы, которые позволяют провести эксперимент для отыскания уравнения регрессии, содержащего только первые степени факторов и их произведения:

$$\hat{y} = b_0 + \sum_{i=1}^k b_i x_i + \sum_{\substack{i,u=1 \\ i \neq u}}^k b_{iu} x_i x_u + \sum_{\substack{i,j,u=1 \\ i \neq j \neq u}}^k b_{iju} x_i x_j x_u + \dots \quad (6.6)$$

Планы второго порядка позволяют провести эксперимент для отыскания уравнения регрессии, содержащего и вторые степени факторов:

$$\hat{y} = b_0 + \sum_{i=1}^k b_i x_i + \sum_{i=1}^k b_{ii} x_i^2 + \sum_{\substack{i,u=1 \\ i \neq u}}^k b_{iu} x_i x_u + \dots \quad (6.7)$$

Нахождение уравнения регрессии методом планирования экспериментов состоит из следующих этапов:

- выбор основных факторов и их уровней;
- планирование и проведение собственно эксперимента;
- определение коэффициентов уравнения регрессии;
- статистический анализ результатов эксперимента.

6.3. Планирование первого порядка

На первой стадии исследования обычно принимают полином первой степени. Так, для трехфакторной задачи теоретическое уравнение регрессии имеет вид

$$y = \beta_0 + \sum_{i=1}^3 \beta_i x_i + \sum_{\substack{i,u=1 \\ i \neq u}}^3 \beta_{iu} x_i x_u + \beta_{123} x_1 x_2 x_3. \quad (6.8)$$

Уравнение регрессии, получаемое на основании результатов эксперимента, в отличие от приведенного теоретического уравнения, имеет вид

$$\hat{y} = b_0 + \sum_{i=1}^3 b_i x_i + \sum_{\substack{i,u=1 \\ i \neq u}}^3 b_{iu} x_i x_u + b_{123} x_1 x_2 x_3, \quad (6.9)$$

где коэффициенты регрессии $b_0, b_1, \dots, b_3, \dots, b_{123}$ являются оценками для теоретических коэффициентов регрессии, т.е.

$$b_i \rightarrow \beta_i, \quad b_{iu} \rightarrow \beta_{iu}, \quad b_{123} \rightarrow \beta_{123}.$$

Члены, содержащие произведения $x_1 x_2, x_2 x_3$ и т.д., называют членами, отражающими попарное взаимодействие факторов, члены вида $x_1 x_2 x_3$ — членами тройного взаимодействия.

6.3.1. Выбор основных факторов и их уровней

В качестве факторов можно выбирать только контролируемые и управляемые переменные, т.е. такие, которые исследователь может поддерживать постоянными в течение каждого опыта на заданном уровне. В число факторов должны быть включены параметры процесса, оказывающие наиболее сильное влияние на функцию отклика. Необходимо заметить, что, несмотря на всю заманчивость и очевидные преимущества активного спланированного эксперимента перед пассивным, в его применении имеется целый ряд трудностей, связанных с определенными ограничениями на его реализацию. Важнейшим условием применимости этого подхода является управляемость процессов по каждому из выбранных факторов, т.е. возможность независимого изменения каждого из этих факторов и поддержания его на заданном уровне в период проведения опытов.

Для каждого фактора необходимо указать тот интервал изменения параметров, в пределах которого ставится исследование. Для этого на основе априорной информации устанавливаются ориентировочные значения факторов $x_{10},$

$x_{20}, \dots, x_{i0}, \dots, x_{k0}$. Этой комбинации значений факторов соответствует точка в многомерном факторном пространстве, которая принимается за исходную точку. Координаты этой точки принимаются за основной (нулевой) уровень.

Интервалом варьирования факторов называется некоторое число (каждое для соответствующего фактора), прибавление которого к основному уровню дает верхний, а вычитание – нижний пределы. Для упрощения записи условий эксперимента и обработки экспериментальных данных масштабы по осям выбираются так, чтобы верхний уровень составлял +1, нижний -1, а основной – 0.

Для факторов с непрерывной областью определения это достигается с помощью преобразования (кодирования) факторов:

$$X_i = \frac{x_i - x_{i0}}{\Delta x_i}. \quad (6.10)$$

В теории планирования экспериментов показано, что минимально необходимое число уровней факторов на единицу больше порядка уравнения.

6.3.2. Планирование эксперимента

Рассмотрим сначала частный случай, когда функция отклика линейно зависит от трех независимых факторов. Уравнение регрессии в этом случае имеет вид (6.9), а план эксперимента представлен в табл. 6.3.

Таблица 6.3
Таблица полного факторного эксперимента для трех факторов

Номер опыта	План								Результат y_j
	x_0	x_1	x_2	x_3	$x_1 x_2$	$x_1 x_3$	$x_2 x_3$	$x_1 x_2 x_3$	
1	+1	-1	-1	-1	+1	+1	+1	-1	y_1
2	+1	+1	-1	-1	-1	-1	+1	+1	y_2
3	+1	-1	+1	-1	-1	+1	-1	+1	y_3
4	+1	+1	+1	-1	+1	-1	-1	-1	y_4
5	+1	-1	-1	+1	+1	-1	-1	+1	y_5
6	+1	+1	-1	+1	-1	+1	-1	-1	y_6
7	+1	-1	+1	+1	-1	-1	+1	-1	y_7
8	+1	+1	+1	+1	+1	+1	+1	+1	y_8

Здесь добавлен столбец фиктивной переменной x_0 , нужный для оценки свободного члена b_0 . После реализации плана получают 8 уравнений с 8 неизвестными, их решение и даст оценку всех 8 коэффициентов регрессии $b_0, b_1, \dots, b_3, b_{12}, \dots, b_{123}$.

План, в котором число опытов равно числу определяемых коэффициентов, называется насыщенным.

Заметим, что мы использовали все точки с "крайними" координатами, т.е. ± 1 , или, говоря другими словами, все возможные комбинации выбранных уровней. В самом деле, всех возможных комбинаций $2^k=8$ (k – число факторов), и мы все их использовали. Если эксперименты проводятся только на двух уровнях (при двух значениях факторов) и при этом в процессе эксперимента осуществляются все возможные неповторяющиеся комбинации из k факторов, то постановка опытов по такому плану носит название полного факторного эксперимента (ПФЭ) или 2^k .

Иными словами, полный факторный эксперимент (ПФЭ) — это эксперимент, реализующий все возможные неповторяющиеся комбинации уровней независимых факторов.

Кодированный план геометрически может быть интерпретирован в виде куба, восемь вершин которого представляют собой восемь экспериментальных точек (рис.6.1).

При числе факторов $k=2$ построение матрицы ПФЭ не вызывает затруднений, при увеличении же числа факторов возникает необходимость в некоторых специальных приемах построения матрицы.

Первый прием основан на чередовании знаков. В первом столбце (для x_1) знаки чередуются поочередно. Во втором (для x_2) — через 2, в третьем (для x_3) — через 4 и т.д. по степеням двойки 2^k . Этот подход и использован при составлении плана, представленного в табл.6.3.

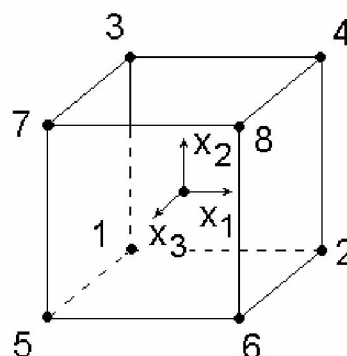


Рис.6.1. Геометрическое изображение ПФЭ

Второй прием основан на последовательном достраивании матрицы. Для этого при добавлении нового фактора необходимо повторить комбинации уровней исходного плана — сначала при значениях нового фактора на верхнем уровне, а затем — на нижнем.

Матрица ПФЭ обладает следующими свойствами:

1) свойство симметричности: алгебраическая сумма элементов вектор-столбца каждого фактора равна нулю (за исключением столбца, соответствующего свободному члену):

$$\sum_{j=1}^n X_{ij} = 0, \quad (6.11)$$

где i — номер фактора; j — номер опыта;

2) свойство нормирования: сумма квадратов элементов каждого столбца равна числу опытов:

$$\sum_{j=1}^n X_{ij}^2 = n; \quad (6.12)$$

3) свойство ортогональности: скалярное произведение всех вектор-столбцов (сумма почленных произведений элементов любых двух вектор-столбцов матрицы) равно нулю:

$$\sum_{j=1}^n X_{ij} X_{uj} = 0, \quad i \neq u. \quad (6.13)$$

Планы, для которых выполняется свойство 3, называют ортогональными. Благодаря этому свойству резко уменьшаются трудности, связанные с расчетом коэффициентов уравнения регрессии.

Поскольку результаты наблюдений отклика носят случайный характер, приходится в каждой точке плана проводить не один, а m^* параллельных опытов (обычно $m^*=2\div 4$), осреднение результатов которых, как уже отмечалось, дает возможность уменьшить погрешности оценки истинного значения отклика в $\sqrt{m^*}$ раз.

В каждой серии экспериментов их последовательность рандомизируется, т.е. с помощью таблиц случайных чисел определяется случайная последова-

тельность реализации экспериментов. Рандомизация дает возможность свести эффект некоторого случайного фактора к случайной погрешности. Это позволяет в определенной степени исключить предвзятость и субъективизм исследователя.

6.3.3. Определение коэффициентов уравнения регрессии

Воспользуемся свойствами ПФЭ для определения коэффициентов уравнения регрессии методом наименьших квадратов $\hat{y} = b_0 + b_1x_1 + b_2x_2$.

$$\Phi = \sum_{j=1}^n (y_j - \hat{y}_j)^2 \rightarrow \min_{b_i};$$

$$\frac{\partial \Phi}{\partial b_1} = 2 \sum_{j=1}^n (y_j - b_0 - b_1X_{1j} - b_2X_{2j})X_{1j} = 0; \quad (6.14)$$

$$\sum_{j=1}^n y_j X_{1j} - b_0 \sum_{j=1}^n X_{1j} - b_1 \sum_{j=1}^n X_{1j}^2 - b_2 \sum_{j=1}^n X_{1j} X_{2j} = 0.$$

Воспользуемся свойствами ПФЭ:

- (симметричности) $b_0 \sum X_{1j} = 0$;
- (нормирования) $b_1 \sum X_{1j}^2 = nb_1$;
- (ортогональности) $b_2 \sum X_{1j} X_{2j} = 0$;

$$b_1 = \frac{\sum_{j=1}^n y_j X_{1j}}{n}; \quad b_2 = \frac{\sum_{j=1}^n y_j X_{2j}}{n}; \quad b_0 = \frac{\sum_{j=1}^n y_j X_{0j}}{n}. \quad (6.15)$$

Следовательно, любые коэффициенты уравнения регрессии определяются скалярным произведением столбца y на соответствующий столбец X .

Можно показать, что аналогичным образом определяются коэффициенты, если в уравнении регрессии (6.6) учитываются линейные взаимодействия (двойные, тройные):

$$b_{12} = \frac{\sum_{j=1}^n y_j(X_1X_2)_j}{n}; \quad b_{123} = \frac{\sum_{j=1}^n y_j(X_1X_2X_3)_j}{n} \text{ и т.д.} \quad (6.16)$$

Следует обратить особое внимание на то, что все линейные коэффициенты независимы, так как в формулы для их расчета (6.15), (6.16) входят свои одноименные переменные. Поэтому каждый коэффициент характеризует роль соответствующей переменной в процессе или силу влияния факторов. Чем больше численная величина коэффициента, тем большее влияние оказывает этот фактор. Если коэффициент имеет знак плюс, то с увеличением значения фактора отклик увеличивается, а если минус – уменьшается.

В результате определения уравнения регрессии может получиться так, что один (или несколько) коэффициентов не очень большие и окажутся незначимыми. Факторы, имеющие коэффициенты, незначимо отличающиеся от нуля, могут быть выведены из состава уравнения, так как их влияние на параметры отклика будет отнесено к ошибке эксперимента. Учитывая ортогональность плана, оставшиеся коэффициенты уравнения регрессии можно не пересчитывать. При отсутствии ортогональности плана эксперимента все коэффициенты необходимо пересчитывать заново.

6.3.4. Статистический анализ результатов эксперимента

Планирование эксперимента исходит из статистического характера зависимостей, поэтому полученные уравнения подвергаются тщательному статистическому анализу с целью извлечь из результатов эксперимента максимум информации и убедиться в достоверности полученной зависимости и ее точности. Процедура проверки значимости коэффициентов уравнения регрессии и его адекватности принципиально не отличается от описания, данного в параграфах 4.5.1 и 4.5.2, поэтому остановимся только на отдельных моментах. Как уже отмечалось ранее, каждый эксперимент несет в себе какую-то погрешность, для повышения надежности результатов производятся для каждой строки таблицы планирования повторения опытов m^* раз.

Построчные (выборочные) дисперсии подсчитываются по формуле

$$S_j^2 = \frac{\sum_{i=1}^{m^*} (y_{ji} - \bar{y}_j)^2}{m^* - 1}, \quad (6.17)$$

где $\bar{y}_j = \frac{\sum_{i=1}^{m^*} y_{ji}}{m^*}$ – средний отклик по m^* опытам в точке с номером j .

Дисперсия воспроизводимости отклика $S_{\text{ВОСП}}^2$ есть среднеарифметическое дисперсий всех n различных вариантов опытов:

$$S_{\text{восп}}^2 = \frac{\sum_{j=1}^n S_j^2}{n} = \frac{\sum_{j=1}^n \sum_{i=1}^{m^*} (y_{ji} - \bar{y}_j)^2}{n(m^* - 1)}. \quad (6.18)$$

Прежде чем производить объединение дисперсий, следует убедиться в их однородности. Проверка производится с помощью критерия Фишера или Кохрена (см. гл. 3). Для оценки значимости коэффициентов прежде всего находят дисперсию коэффициентов регрессии. Учитывая свойства 1-3 плана, представленного в табл. 6.3, из выражений (4.27) и (4.27а) при одинаковом дублировании опытов по точкам с числом повторных опытов m^* получим

$$S_b^2 = \frac{S_{\text{ВОСП}}^2}{m^* n}, \quad (6.19)$$

а при отсутствии дублирования будем иметь

$$S_b^2 = \frac{S_{\text{ВОСП}}^2}{n}. \quad (6.19a)$$

Следовательно, все коэффициенты уравнения регрессии ПФЭ имеют одинаковую точность (дисперсию). В этом заключается принципиальное отличие коэффициентов уравнения регрессии, полученных по плану табл.6.3, от коэффициентов уравнений, полученных пассивным экспериментом (см. параграф 4.5.2). Планы, по результатам которых коэффициенты уравнения регрессии определяются с одинаковой дисперсией, называются ротатабельными. В связи с этим план, представленный в табл.6.3, является не только ортогональным, но ротатабельным. В дальнейшем проверка значимости каждого коэффициента производится с использованием t -критерия Стьюдента (см. гл.4). Статистически

незначимые коэффициенты исключаются из уравнения, а остальные коэффициенты при этом не пересчитываются. После этого уравнение регрессии составляется в виде уравнения связи выходного параметра y и переменных X_i , включающего только значимые коэффициенты.

После вычисления коэффициентов уравнения следует прежде всего проверить его пригодность или адекватность. Для этого достаточно оценить отклонение выходной величины \hat{y} , предсказанной уравнением регрессии, от результатов эксперимента y в различных точках.

Рассеяние результатов эксперимента относительно уравнения регрессии, аппроксимирующего искомую зависимость, можно, как уже было показано ранее, охарактеризовать с помощью дисперсии адекватности, оценка которой, справедливая при одинаковом числе дублирующих опытов, находится по формуле

$$S_{ад}^2 = \frac{m * \sum_{j=1}^n (\bar{y}_j - \hat{y}_j)^2}{n - l}. \quad (6.20)$$

Здесь n – число опытов (вариантов); $l=k+1$, где k – число членов в уравнении регрессии.

Проверка адекватности состоит в выяснении соотношения между дисперсией адекватности $S_{ад}^2$ и дисперсией воспроизводимости $S_{восп}^2$ и проводится с помощью F-критерия Фишера, который в данном случае рассчитывается как

$$F = \frac{S_{ад}^2}{S_{восп}^2}. \quad (6.21)$$

Если вычисленное значение критерия меньше теоретического $F_{\alpha; m_1; m_2}$ для соответствующих степеней свободы $m_1=n-l$, $m_2=n(m^*-1)$, при заданном уровне значимости α , то описание свойств объекта уравнением регрессии признается адекватным объекту. Адекватность модели может быть достигнута уменьшением интервала варьирования факторов, а если это не дает результата, то переходом к плану второго порядка.

6.3.5. Дробный факторный эксперимент

Во многих практических задачах взаимодействия второго и высших порядков отсутствуют или пренебрежимо малы. Кроме того, на первых этапах исследования часто необходимо получить в первом приближении лишь линейную аппроксимацию изучаемого уравнения связи при минимальном числе экспериментов. Так, для трех факторов вместо уравнения (6.9) достаточно рассмотреть уравнение вида

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \quad (6.22)$$

и определить только четыре коэффициента. Поэтому использование ПФЭ для определения коэффициентов только при линейных членах неэффективно из-за реализации большого числа опытов, особенно при большом числе факторов k .

Если при решении задачи можно ограничиться линейным приближением, то в ПФЭ оказывается много "лишних" опытов. Так, для трех факторов достаточно 4 опыта, а в ПФЭ их 8. Следовательно, есть четыре "лишних". Результаты этих "лишних" опытов могут быть использованы двояко: во-первых, с их помощью можно получить более точные оценки коэффициентов регрессии; во-вторых, их можно использовать для проверки адекватности модели. Однако при 7 факторах ПФЭ содержит $2^7=128$ опытов, а для линейного уравнения требуется всего 8. Таким образом, остается 120 лишних и, конечно, нет необходимости их все реализовать, а достаточно лишь несколько из них использовать для проверки адекватности и уточнения оценок.

Другими словами, ПФЭ обладает большой избыточностью опытов. В связи с этим возникает вопрос: "Нельзя ли сократить число опытов, необходимых для определения коэффициентов регрессии?"

Так, для определения коэффициентов уравнения (6.22) достаточно ограничиться четырьмя опытами, если в ПФЭ 2^3 использовать x_1x_2 в качестве плана для x_3 , тогда матрица планирования эксперимента примет вид, представленный в табл.6.4.

Дробный факторный эксперимент

Номер опыта	План				Результат y_i
	x_0	x_1	x_2	$x_3 = x_1 x_2$	
1	+1	-1	-1	+1	y_1
2	+1	+1	-1	-1	y_2
3	+1	-1	+1	-1	y_3
4	+1	+1	+1	+1	y_4

Заметим, что мы использовали не все точки с "крайними" координатами, т.е. ± 1 , или, говоря другими словами, не все возможные комбинации выбранных уровней. На самом деле всех возможных комбинаций $2^3=8$, мы же использовали из них только 4. Такой сокращенный план носит название дробного факторного эксперимента (ДФЭ).

Следует подчеркнуть, что формальное приравнение произведения факторов фактору, не входящему в это произведение, является основополагающей идеей метода ДФЭ. В данном случае используется только половина ПФЭ 2^3 , поэтому план, представленный в табл.6.4, называется полурепликой от ПФЭ 2^3 . После реализации плана получают 4 уравнения с 4 неизвестными, их решение и даст оценку всех четырех коэффициентов регрессии b_i . Например, матрица из 8 опытов для четырехфакторного планирования будет полурепликой от ПФЭ 2^4 , а для пятифакторного планирования — четвертьрепликой от 2^5 .

Для того чтобы дробная реплика представляла собой ортогональный план, в качестве реплики следует брать ближайший полный факторный эксперимент. При этом число опытов должно быть не менее числа искомых коэффициентов.

Если коэффициенты регрессии при парных произведениях не равны нулю, то найденные коэффициенты b_i будут смешанными оценками их теоретических коэффициентов β_i . На практике обычно не удается априорно постулировать равенство нулю эффектов взаимодействия, однако часто имеются основания полагать, что некоторые из них малы по сравнению с линейными эффектами. Операцию смешивания оценок принято условно записывать в виде выражений

$$b_1 \rightarrow \beta_1 + \beta_{23}; b_2 \rightarrow \beta_2 + \beta_{13}; b_3 \rightarrow \beta_3 + \beta_{12}, \quad (6.23)$$

где β – математическое ожидание для соответствующего коэффициента.

Эти генерирующие коэффициенты не могут быть отдельно оценены по плану, включающему всего четыре опыта, так как в этом случае неразличимы столбцы для линейных членов и парных произведений. Если, например, в дополнение к столбцам, приведенным в табл. 6.4, вычислить еще столбцы для произведения x_1x_3 , то увидим, что элементы этого столбца в точности равны элементам столбца x_2 . Таким образом, сокращение числа опытов приводит к получению смешанных оценок для коэффициентов.

Для того чтобы определить, какие коэффициенты смешаны, удобно пользоваться следующим приемом: подставив x_3 на место x_1x_2 , получим соотношение $x_3=x_1x_2$, называемое генерирующим соотношением.

Умножив обе части генерирующего соотношения на x_3 , получим

$$X_3^2 = X_1X_2X_3 = 1, \text{ т.е. } X_1X_2X_3 = 1. \quad (6.24)$$

Это произведение носит название определяющего контраста.

Умножив поочередно определяющий контраст на x_1, x_2, x_3 , находим

$$X_1 = X_1^2X_2X_3 = X_2X_3; X_2 = X_1X_3; X_3 = X_1X_2. \quad (6.25)$$

Полученным соотношениям соответствует система смешанных оценок, т.е. β_1 смешана с β_{23} , β_2 – с β_{13} , а β_3 – с β_{12} .

Таким образом, при использовании ДФЭ необходимо иметь четкое представление о так называемой разрешающей способности дробных реплик, т.е. определить заранее, какие коэффициенты являются несмешанными оценками для соответствующих коэффициентов. Тогда в зависимости от постановки задачи подбирается дробная реплика, с помощью которой можно извлечь максимальную информацию из эксперимента.

Например, в задаче с четырьмя факторами ($k=4$) в качестве генерирующего соотношения можно взять $x_4=x_1x_2x_3$ или любой из эффектов двойного взаимодействия, например $x_4=x_1x_2$. Таблица планирования такого эксперимента представлена в табл.6.5.

Таблица 6.5

Планирование ДФЭ

Номер опыта	План				Генерирующие соотношения	
	X ₀	X ₁	X ₂	X ₃	X ₄ =X ₁ X ₂ X ₃	X ₄ =X ₁ X ₂
1	+1	-1	-1	-1	-1	+1
2	+1	+1	-1	-1	+1	-1
3	+1	-1	+1	-1	+1	-1
4	+1	+1	+1	-1	-1	+1
5	+1	-1	-1	+1	+1	+1
6	+1	+1	-1	+1	-1	-1
7	+1	-1	+1	+1	-1	-1
8	+1	+1	+1	+1	+1	+1

В первом случае определяющий контраст $X_4^2 = X_1 X_2 X_3 X_4 = 1$. Получим оценку совместных оценок:

$$X_1 = X_2 X_3 X_4; b_1 \rightarrow \beta_1 + \beta_{234};$$

$$X_2 = X_1 X_3 X_4; b_2 \rightarrow \beta_2 + \beta_{134};$$

$$X_3 = X_1 X_2 X_4; b_3 \rightarrow \beta_3 + \beta_{124};$$

$$X_4 = X_1 X_2 X_3; b_4 \rightarrow \beta_4 + \beta_{123};$$

$$X_1 X_4 = X_2 X_3; b_{14} \rightarrow \beta_{14} + \beta_{23};$$

$$X_1 X_2 = X_3 X_4; b_{12} \rightarrow \beta_{12} + \beta_{34};$$

$$X_1 X_3 = X_2 X_4; b_{13} \rightarrow \beta_{13} + \beta_{24}.$$

В реальных задачах тройные взаимодействия бывают равными нулю значительно чаще, чем двойные. Значит, если по физическому смыслу задачи нас более всего интересуют оценки для линейных эффектов, следует использовать генерирующее соотношение $X_4 = X_1 X_2 X_3$.

Во втором случае определяющий контраст выражается соотношением $X_4^2 = X_1 X_2 X_4 = 1$; $X_1 X_2 X_4 = 1$.

При этом получим следующую систему оценок:

$$X_1 = X_2 X_4; b_1 \rightarrow \beta_1 + \beta_{24};$$

$$X_2 = X_1 X_4; b_2 \rightarrow \beta_2 + \beta_{14};$$

$$X_3 = X_1X_2X_3X_4; b_3 \rightarrow \beta_3 + \beta_{1234};$$

$$X_4 = X_1X_2; b_4 \rightarrow \beta_4 + \beta_{12};$$

$$X_1X_3 = X_2X_3X_4; b_{13} \rightarrow \beta_{13} + \beta_{234};$$

$$X_2X_3 = X_1X_3X_4; b_{23} \rightarrow \beta_{23} + \beta_{134};$$

$$X_3X_4 = X_1X_2X_3; b_{34} \rightarrow \beta_{34} + \beta_{123}.$$

Следовательно, дробную реплику с генерирующим соотношением $X_4=X_1X_2$ имеет смысл использовать, если нас более всего интересуют коэффициенты β_{12} , β_{23} , β_{34} .

Дробную реплику, в которой P линейных эффектов приравнены к эффектам взаимодействия, обозначают 2^{k-P} .

Таким образом, планы первого порядка, оптимальные двухуровневые планы ПФЭ 2^k и ДФЭ 2^{k-P} имеют следующие преимущества:

- 1 – планы ортогональны, поэтому все вычисления просты;
- 2 – все коэффициенты определяются независимо один от другого;
- 3 – каждый коэффициент определяется по результатам всех n опытов;
- 4 – все коэффициенты регрессии определяются с одинаковой дисперсией, т.е. эти планы обладают и свойством ротатабельности.

6.3.6. Разработка математической модели гидравлического режима методической печи

В качестве примера рассмотрим разработку математической модели гидравлического режима четырехзонной методической печи с использованием теории планирования эксперимента. При планировании опытов используем методику проведения дробного факторного эксперимента (ДФЭ) первого порядка с двухуровневым варьированием факторов.

Перед разработкой плана эксперимента на основе априорной информации были выявлены факторы, влияющие на величину давления в томильной зоне печи. К числу таких факторов относятся расходы топлива на каждую зону нагрева и угол поворота дымового клапана. Расходы воздуха на каждую зону в

качестве факторов не фигурировали, поскольку схема управления горением топлива автоматически меняет расход воздуха при изменении расхода газа.

Обозначим факторы: X_1 — расход газа в томильной зоне, $\text{м}^3/\text{ч}$; X_2 — расход газа во второй сварочной зоне, $\text{м}^3/\text{ч}$; X_3 — расход газа в первой сварочной зоне, $\text{м}^3/\text{ч}$; X_4 — расход газа в нижней сварочной зоне, $\text{м}^3/\text{ч}$; X_5 — положение дымового клапана, % хода исполнительного механизма (рис.6.2).

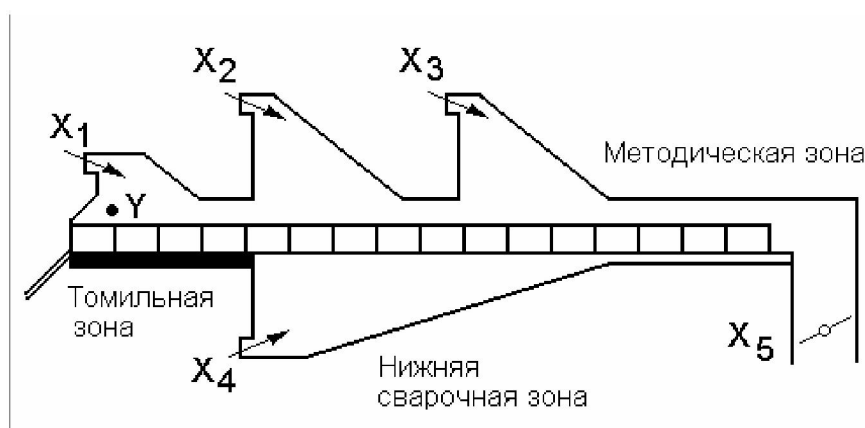


Рис.6.2. Положение факторов (X_1, \dots, X_5) и отклика (Y) при проведении исследования на методической печи

Реализация ПФЭ в этом случае при варьировании всех факторов на двух уровнях потребовала бы постановки $2^5=32$ опытов.

Будем предполагать, что эффекты взаимодействия факторов в исследуемом объекте маловероятны и пренебрежимо малы. Воспользуемся $1/4$ репликой ПФЭ, т.е.ДФЭ типа 2^{5-2} , где формально 2 фактора заменены соответствующими произведениями остальных факторов ($x_4=x_1 x_2$, $x_5=x_1 x_2 x_3$). Это позволит сократить число опытов до $2^3=8$. Уровни варьирования факторов представлены в табл. 6.6.

Таблица 6.6

Уровни варьирования факторов

Уровни факторов	Факторы				
	X_1 , $\text{м}^3/\text{ч}$	X_2 , $\text{м}^3/\text{ч}$	X_3 , $\text{м}^3/\text{ч}$	X_4 , $\text{м}^3/\text{ч}$	X_5 , % хода ИМ
Основной (нулевой)	5250	3900	2650	1100	74
Нижний	4000	3100	1750	700	50
Верхний	6500	4700	3500	1500	98
Интервал варьирования	1250	800	900	400	24

В табл. 6.7 приведены матрица планирования ДФЭ 2^{5-2} и результаты эксперимента — значения выходной переменной (давления в томильной зоне методической печи).

Для обработки результатов эксперимента используем методику, изложенную ранее в параграфе 4.5.

1. Расчет построчных средних:

$$\bar{y}_j = \frac{y_{j1} + y_{j2} + \dots + y_{jm^*}}{m^*},$$

где m^* — число повторных опытов ($m^*=2$). Например,

$$\bar{y}_1 = \frac{(-2,6) + (-2,5)}{2} = -2,55.$$

Результаты расчета представлены в табл. 6.7.

Таблица 6.7

Матрица ДФЭ 2^{5-2} с двумя параллельными опытами

Факторы (кодированные значения)						Переменная состояния (отклик), кПа				Построчная дисперсия S_j^2
						Опы т 1	Опы т 2	Сред- нее	Мо- дель	
X_0	X_1	X_2	X_3	X_4	X_5	y_1	y_2	\bar{y}_j	\hat{y}_j	
1	1	1	1	1	1	-2,5	-2,6	-2,55	-2,41	0,005
1	-1	1	1	-1	-1	2,2	2,3	2,25	2,26	0,005
1	1	-1	1	-1	-1	5,1	4,7	4,90	4,74	0,080
1	-1	-1	1	1	1	-1,1	0,5	-0,30	0,08	1,280
1	1	1	-1	1	-1	2,1	2,3	2,20	2,26	0,020
1	-1	1	-1	-1	1	-2,0	-2,4	-2,20	-2,41	0,080
1	1	-1	-1	-1	1	0,0	0,8	0,40	0,08	0,320
1	-1	-1	-1	1	-1	4,2	5,1	4,65	4,74	0,405

2. Определение построчных (выборочных) дисперсий:

$$S_j^2 = \frac{\sum_{i=1}^{m^*} (y_{ji} - \bar{y}_j)^2}{m^* - 1}; S_1^2 = \frac{(-2,5 - (-2,55))^2 + (-2,6 - (-2,55))^2}{2 - 1} = 0,005.$$

Аналогично $S_2^2=0,005$; $S_3^2=0,08$; $S_4^2=1,28$; $S_5^2=0,02$; $S_6^2=0,08$; $S_7^2=0,32$; $S_8^2=0,405$. Сумма построчных (выборочных) дисперсий:

$$S_{\Sigma}^2 = 0,005 + 0,005 + 0,08 + 1,28 + 0,02 + 0,08 + 0,32 + 0,405 = 2,195.$$

3. Определение однородности дисперсий по критерию Кохрена:

$$G_{\text{эксп}} = \frac{S_{j\text{max}}^2}{S_{\Sigma}^2} = \frac{1,28}{2,195} = 0,5831.$$

Далее по табл. П.9 находим $G_{\alpha;m;n}$. Для $\alpha=0,05$, $m=m^*-1=2-1$ и $n=8$ значение $G_{0,05;1;8}=0,6798$. Поскольку $G_{\text{эксп}} < G_{\text{теор}}$, то дисперсии однородны.

4. Определение коэффициентов в уравнении регрессии:

$$b_0 = \frac{\sum_{j=1}^n \bar{y}_j \cdot x_{0j}}{n} = \frac{-2,55 + 2,25 + 4,9 - 0,3 + 2,2 - 2,2 + 0,4 + 4,65}{8} = 1,169;$$

$$b_1 = \frac{\sum_{j=1}^n \bar{y}_j \cdot x_{1j}}{n} = \frac{-2,55 - 2,25 + 4,9 + 0,3 + 2,2 + 2,2 + 0,4 - 4,65}{8} = 0,069;$$

$$b_2 = \frac{\sum_{j=1}^n \bar{y}_j \cdot x_{2j}}{n} = \frac{-2,55 + 2,25 - 4,9 + 0,3 + 2,2 - 2,2 - 0,4 - 4,65}{8} = -1,244;$$

$$b_3 = \frac{\sum_{j=1}^n \bar{y}_j \cdot x_{3j}}{n} = \frac{-2,55 + 2,25 + 4,9 - 0,3 - 2,2 + 2,2 - 0,4 - 4,65}{8} = -0,094;$$

$$b_4 = \frac{\sum_{j=1}^n \bar{y}_j \cdot x_{4j}}{n} = \frac{-2,55 - 2,25 - 4,9 - 0,3 + 2,2 + 2,2 - 0,4 + 4,65}{8} = -0,169;$$

$$b_5 = \frac{\sum_{j=1}^n \bar{y}_j \cdot x_{5j}}{n} = \frac{-2,55 - 2,25 - 4,9 - 0,3 - 2,2 - 2,2 + 0,4 - 4,65}{8} = -2,331.$$

5. Проверка значимости коэффициентов регрессии. Предварительно определим дисперсию воспроизводимости (дисперсию отклика):

$$S_{\text{восп}}^2 = \frac{\sum_{j=1}^n S_j^2}{n} = \frac{S_{\Sigma}^2}{n} = \frac{2,195}{8} = 0,2744.$$

Дисперсия коэффициентов уравнения регрессии

$$S_b^2 = \frac{S_{\text{восп}}^2}{n \cdot m^*} = \frac{0,2744}{8 \cdot 2} = 0,01715; \quad S_b = \sqrt{S_b^2} = 0,131.$$

Находим значение доверительного интервала для коэффициентов регрессии:

$$\Delta b_j = t_{\alpha; m} \cdot S_b.$$

Здесь $m = n(m^* - 1) = 8(2 - 1) = 8$, тогда теоретическое значение критерия Стьюдента $t_{0,05;8} = 2,31$ (можно рассчитать, используя функцию электронных таблиц Microsoft Excel **СТЮДРАСПОБР**(0,05;8)=2,31), откуда $\Delta b_j = 2,31 \cdot 0,131 = 0,303$. Из сопоставления доверительного интервала Δb_j с абсолютными значениями коэффициентов модели следует, что $|b_1| = 0,069 < 0,303$; $|b_3| = 0,094 < 0,303$ и $|b_4| = 0,169 < 0,303$. Эти коэффициенты оказались незначимы, а остальные значимы. Таким образом, окончательное уравнение регрессии запишется в виде

$$\hat{y} = 1,169 - 1,244X_2 - 2,331X_5.$$

Результаты расчета выходных параметров по уравнению полученной модели \hat{y}_i занесены в табл. 6.7.

6. Проверка адекватности полученной модели. Предварительно определим дисперсию адекватности:

$$S_{ад}^2 = \frac{m^* \sum_{i=1}^n (\bar{y}_i - \hat{y}_i)^2}{n - l}.$$

В нашем случае $m^* = 2$; $n = 8$; $l = 3$, и в результате имеем

$$S_{ад}^2 = \frac{2}{8 - 3} [(-2,55 + 2,41)^2 + (2,25 - 2,66)^2 + (4,9 - 4,74)^2 + (-0,3 - 0,08)^2 + (2,2 - 2,26)^2 + (-2,2 + 2,41)^2 + (0,4 - 0,08)^2 + (4,65 - 4,74)^2] = 0,1386.$$

С учетом ранее найденной выборочной дисперсии $S_{\Sigma}^2 = 2,195$ определяем дисперсию воспроизводимости:

$$S_{восп}^2 = \frac{S_{\Sigma}^2}{n} = \frac{2,195}{8} = 0,274.$$

Экспериментальное значение критерия Фишера следующее:

$$F_{эксп} = \frac{S_{ад}^2}{S_{восп}^2} = \frac{0,1386}{0,2744} = 0,505.$$

Теоретическое значение критерия Фишера $F_{\alpha; m_1; m_2}$ при $\alpha = 0,05$ можно определить по справочнику [11], табл. П.4 или с помощью встроенной функции элек-

тронных таблиц Microsoft Excel **ФРАСПОБР**. Для $m_1=(n-1)=(8-3)=5$ и $m_2=n(m^*-1)=8(2-1)=8$ значение $F_{0,05;5;8}=3,69$ (**ФРАСПОБР**(0,05;5;8)=3,69). Поскольку $F_{\text{эксп}} < F_{\text{теор}}$, то полученная модель адекватна.

6.4. Планы второго порядка

Описание поверхности отклика полиномами первого порядка часто оказывается недостаточным. Во многих случаях удовлетворительная аппроксимация может быть достигнута, если воспользоваться полиномом второго порядка (6.7).

В этом случае требуется, чтобы каждый фактор варьировался не менее чем на трех уровнях. В этом случае полный факторный эксперимент содержит слишком большое количество опытов, равное 3^k . Так, при $k=3$ их 27, а число коэффициентов $b - 10$, при $k=5$ число опытов 243, а коэффициентов 21. В связи с этим осуществление ПФЭ для планов второго порядка не только сложно, но и нецелесообразно.

Сократить число опытов можно, воспользовавшись так называемым композиционным или последовательным планом, разработанным Боксом и Уилсоном. Так, при двух факторах модель функции отклика $y = f(x_1, x_2)$ второго порядка представляет собой поверхность в виде цилиндра, конуса, эллипса и т.д., описываемую в общем виде уравнением

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2.$$

Для определения такой поверхности необходимо располагать координатами не менее трех ее точек, т.е. факторы x_1 и x_2 должны варьироваться не менее чем на трех уровнях. Поэтому план эксперимента в плоскости факторов x_1 и x_2 на рис.6.3, а не может состоять лишь из опытов 1, 2, 3, 4 ПФЭ 2^2 , располагающихся в вершинах квадрата, как это было для модели первого порядка. К ним должны быть добавлены опыты (звездные точки) 5, 6, 7, 8, расположенные на осях x_1 и x_2 с координатами $(\pm\alpha; 0)$, $(0; \pm\alpha)$ и обязательно опыт 9 в центре квадрата, чтобы по любому направлению (5-9-6), (1-9-4) и т.д. располагались три точки, определяющие кривизну поверхности в этом направлении.

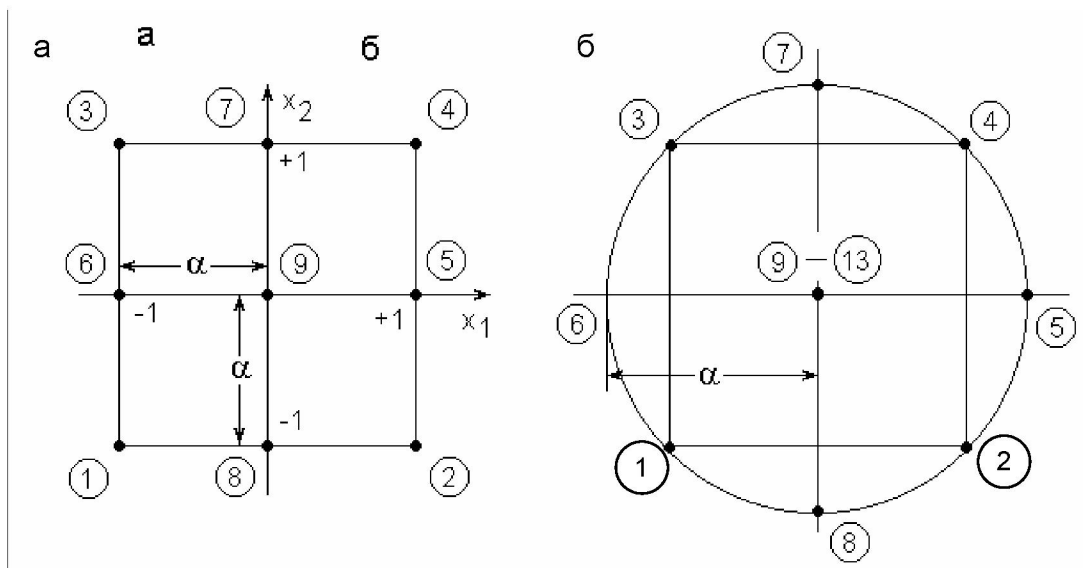


Рис. 6.3. Планы второго порядка при $k=2$:
 а — ортогональный; б — ротатабельный

Таким образом, в общем случае ядро композиционного плана составляет при $k < 5$ ПФЭ 2^k , а при $k \geq 5$ — дробную реплику от него. Если линейное уравнение регрессии оказалось неадекватным, необходимо:

1) добавить $2 \cdot k$ звездных точек, расположенных на координатных осях факторного пространства $(\pm\alpha, 0, 0, \dots, 0)$, $(0, \pm\alpha, 0, \dots, 0)$, ..., $(0, 0, \dots, \pm\alpha)$, где α — звездное плечо, или расстояние до звездной точки;

2) провести n_0 опытов при значениях факторов в центре плана.

При k факторах общее число опытов в матрице композиционного плана составит

$$n = 2^k + 2 \cdot k + n_0 \text{ при } k < 5,$$

$$n = 2^{k-1} + 2 \cdot k + n_0 \text{ при } k \geq 5.$$

При этом величина звездного плеча α и число опытов в центре плана n_0 зависит от выбранного вида композиционного плана.

Композиционный план для $k=2$ и $n_0=1$ представлен в табл.6.8.

Таблица 6.8

Композиционный план второго порядка

Номер	Ф а к т о р ы	Результат
-------	---------------	-----------

опыта		x_0	x_1	x_2	x_1x_2	x_1^2	x_2^2	y_j
Ядро плана	1	+1	-1	-1	+1	+1	+1	y_1
	2	+1	+1	-1	-1	+1	+1	y_2
	3	+1	-1	+1	-1	+1	+1	y_3
	4	+1	+1	+1	+1	+1	+1	y_4
	5	+1	$+\alpha$	0	0	α^2	0	y_5
Звезд- ные точки	6	+1	$-\alpha$	0	0	α^2	0	y_6
	7	+1	0	$+\alpha$	0	0	α^2	y_7
	8	+1	0	$-\alpha$	0	0	α^2	y_8
Центр плана	9	+1	0	0	0	0	0	y_9

Аналогичным образом составляются планы и для большего числа факторов.

6.4.1. Ортогональные планы второго порядка

В общем виде план, представленный в табл.6.8, неортогонален, так как

$$\sum_{j=1}^n x_{0j}x_{ij}^2 \neq 0; \sum_{j=1}^n x_{ij}^2x_{uj}^2 \neq 0, i \neq u. \quad (6.26)$$

Приведем его к ортогональному виду, для чего введем новые переменные (преобразования для квадратичных эффектов):

$$x'_{ij} = x_{ij}^2 - \frac{\sum_{j=1}^n x_{ij}^2}{n} = x_{ij}^2 - \bar{x}_i^2.$$

$$\text{При этом } \sum_{j=1}^n x_{0j}x'_{ij} = \sum_{j=1}^n (x_{ij}^2 - \bar{x}_i^2) = \sum_{j=1}^n x_{ij}^2 - n\bar{x}_i^2 = 0.$$

Тогда уравнение регрессии будет записано как

$$\hat{y} = b_0' + \sum_{i=1}^k b_i x_i + \sum_{i,u=1}^k b_{iu} x_i x_u + \sum_{i=1}^k b_{ii}' x_i'.$$

Композиционные планы легко привести к ортогональным, выбирая звездное плечо α . В табл. 6.9 приведено значение α для различного числа факторов k и числа опытов в центре плана n_0 .

Таблица 6.9

Значения звездных плеч в ортогональных планах второго порядка

Число опытов в центре плана n_0	Звездное плечо α при различном числе факторов k			
	$k=2$	$k=3$	$k=4$	$k=5^*$
1	1,000	1,215	1,414	1,546
2	1,077	1,285	1,471	1,606
3	1,148	1,353	1,546	1,664
4	1,214	1,414	1,606	1,718
5	1,267	1,471	1,664	1,772
6	1,320	1,525	1,718	1,819
7	1,369	1,575	1,772	1,868
8	1,414	1,623	1,819	1,913
9	1,454	1,668	1,868	1,957
10	1,498	1,711	1,913	2,000

*) В ядре полуреплики

В частности, ортогональный план второго порядка для $k=2$ и $n_0=1$ представлен в табл. 6.10, а его геометрическая интерпретация – на рис. 6.3, а.

Представленный на рис.6.3, а и в табл. 6.10 прямоугольный (квадратный) план эксперимента для модели второго порядка работоспособен, хотя и несколько избыточен (9 опытов для определения 6 коэффициентов). Благодаря трем избыточным опытам, он позволяет усреднить случайные погрешности и оценить их характер.

Таблица 6.10

Ортогональный план второго порядка

Номер опыта	Ф а к т о р ы						Резуль- тат y_j	
	x_0	x_1	x_2	x_1x_2	x_1'	x_2'		
Ядро плана	1	+1	-1	-1	+1	+1/3	+1/3	y_1
2	+1	+1	-1	-1	+1/3	+1/3	+1/3	y_2
3	+1	-1	+1	-1	+1/3	+1/3	+1/3	y_3
4	+1	+1	+1	+1	+1/3	+1/3	+1/3	y_4
Звезд- ные точки	5	+1	$\alpha=+1$	0	0	+1/3	-2/3	y_5
	6	+1	$\alpha=-1$	0	0	+1/3	-2/3	y_6
	7	+1	0	$\alpha=+1$	0	-2/3	+1/3	y_7
	8	+1	0	$\alpha=-1$	0	-2/3	+1/3	y_8
Центр плана	9	+1	0	0	0	-2/3	-2/3	y_9

$$\text{В этой таблице } x'_{ij} = x_{ij}^2 - \frac{\sum_{j=1}^9 x_{ij}^2}{9} = x_{ij}^2 - \frac{2}{3}.$$

В силу ортогональности матрицы планирования все коэффициенты

$$b_i = \frac{\sum_{j=1}^n x_{ij} y_j}{\sum_{j=1}^n x_{ij}^2}; \quad b_{ii}' = \frac{\sum_{j=1}^n x_{ij}' y_j}{\sum_{j=1}^n x_{ij}'^2}; \quad b_{iu} = \frac{183 \sum_{j=1}^n (x_{ij} x_{iu}) y_j}{\sum_{j=1}^n (x_{ij} x_{iu})^2}. \quad (6.27)$$

уравнения регрессии b определяются независимо один от другого по формулам
Здесь i – номер столбца в матрице планирования; j – номер строки; суммы в знаменателях различны для линейных, квадратичных эффектов и взаимодействий.

Дисперсии коэффициентов уравнения регрессии следующие:

$$S_{b_i}^2 = S^2_{\text{восп}} / \sum_{j=1}^n x_{ij}^2; \quad S_{b_{ii}}^2 = S^2_{\text{восп}} / \sum_{j=1}^n x'_{ij}{}^2; \quad S_{b_{iu}}^2 = S^2_{\text{восп}} / \sum_{j=1}^n (x_{ij}x_{uj})^2. \quad (6.28)$$

Следует особо отметить, что коэффициенты уравнения регрессии, получаемые с помощью ортогональных планов второго порядка, определяются с разной точностью (см. уравнение (6.28)), в то время как ортогональные планы первого порядка обеспечивают одинаковую точность коэффициентов, т.е. план, представленный в табл. 6.10, являющийся ортогональным и обеспечивающий независимость определения коэффициентов b , не является ротатабельным.

В результате расчетов по матрице с преобразованными столбцами для квадратичных эффектов получим уравнение регрессии в виде

$$\hat{y} = b_0' + \sum_{i=1}^k b_i x_i + \sum_{i,u=1}^k b_{iu} x_i x_u + \sum_{i=1}^k b_{ii}' (x_i^2 - \bar{x}_i^2). \quad (6.29)$$

Для преобразования к обычной форме записи следует перейти от коэффициента b_0' к коэффициенту b_0 , используя выражение

$$b_0 = b_0' - \sum_{i=1}^k b_{ii}' \bar{x}_i^2. \quad (6.30)$$

При этом дисперсия этого коэффициента рассчитывается по следующему соотношению:

$$S_{b_0}^2 = S_{b_0'}^2 + \sum_{i=1}^k \bar{x}_i^2 \cdot S_{b_{ii}'}^2. \quad (6.31)$$

В дальнейшем, зная дисперсию воспроизводимости, проверяют значимость коэффициентов и адекватность уравнения:

$$\hat{y} = b_0 + \sum_{i=1}^k b_i x_i + \sum_{i,u=1}^k b_{iu} x_i x_u + \sum_{i=1}^k b_{ii} x_i^2. \quad (6.32)$$

Значимость коэффициентов проверяется по критерию Стьюдента $t_i = |b_i| / S_{b_i}$. Коэффициент значим, если $t_i > t_{\alpha, m}$, где m – число степеней свободы дисперсии воспроизводимости.

Адекватность уравнения проверяется по критерию Фишера $F = S_{ад}^2 / S_{восп}^2$. Уравнение адекватно, если составленное таким образом F-отношение меньше теоретического: $F < F_{\alpha; m_1; m_2}$, где $m_1 = n - l$ – число степеней свободы дисперсии адекватности; m_2 – число степеней свободы дисперсии воспроизводимости; l – число коэффициентов в уравнении регрессии второго порядка, равное числу сочетаний из $k+2$ по 2, т.е.

$$l = \frac{(k+2)(k+1)}{2}. \quad (6.32a)$$

6.4.2. Ротатабельные планы второго порядка

Как мы установили, план второго порядка, представленный в табл.6.10, не обладает свойством ротатабельности. Ротатабельным называют планирование, для которого дисперсия отклика (выходного параметра) \hat{y} , предсказанного уравнением регрессии, постоянна для всех точек, находящихся на равном расстоянии от центра эксперимента. Экспериментатору заранее не известно, где находится та часть поверхности отклика, которая представляет для него особый интерес, поэтому следует стремиться к тому, чтобы количество информации, содержащееся в уравнении регрессии, было одинаково для всех равноотстоящих от центра эксперимента точек. Действительно, удаление от центра точек 5,6,7,8 в $\sqrt{2} = 1,414$ раза меньше, чем удаление точек 1, 2, 3, 4 (см. рис. 6.3, а), и, следовательно коэффициенты уравнения регрессии определяются с различной дисперсией. Бокс и Хантер предложили ротатабельные планы 2-го порядка. Для того чтобы композиционный план был ротатабельным, величину звездного плеча α выбирают из условия

$$\alpha = 2^{\frac{k}{4}} \text{ при } k < 5 \text{ и } \alpha = 2^{\frac{k-1}{4}} \text{ при } k \geq 5 \quad (6.32б)$$

или в общем случае

$$\alpha = 2^{\frac{k-p}{4}},$$

где k – число факторов; p – дробность реплики (для ПФЭ $p=0$, для полуреплики $p=1$, для четвертьреплики $p=2$ и т.д.).

Число точек в центре плана n_0 увеличивают. В табл. 6.11 приведены значения α и n_0 для различного числа независимых факторов.

Таблица 6.11

Значения звездных плеч и числа точек в центре ротатабельных планов

Параметр плана	Значения параметров при числе независимых факторов								
	2	3	4	5	6	6	6	7	7
Ядро плана	2^2	2^3	2^4	2^5	2^{5-1}	2^6	2^{6-1}	2^7	2^{7-1}
Звездное плечо	1,414	1,682	2,00	2,378	2,00	2,828	2,378	3,333	2,828
Число точек в центре плана n_0	5	6	7	10	6	15	9	21	14

Поясним идею выбора значения звездного плеча α на примере матрицы ротатабельного планирования второго порядка для $k=2$, представленной в табл. 6.12.

Размещение точек этого плана показано на рис. 6.3,б. Для обеспечения ротатабельности точек 5, 6, 7, 8 необходимо удалить их от центра плана на расстояние α в $\sqrt{2}=1,414$ раз большее, чем удаление точек 1, 2, 3, 4 от осей x_2 и x_1 . В результате этого все точки плана (табл. 6.12) оказываются лежащими на окружности. Учитывая существенно большее влияние на функцию отклика случайной ошибки в точке 9, рекомендуется ставить в этой точке плана не один, а несколько дублирующих опытов (в данном случае опыты с 9 до 13) для усреднения полученных результатов и для осуществления статистического анализа результатов всего эксперимента в целом.

Таблица 6.12

Ротатабельный план второго порядка

Номер опыта	Ф а к т о р ы						Результат y_j	
	x_0	x_1	x_2	x_1x_2	x_1^2	x_2^2		
Ядро плана	1	+1	-1	-1	+1	+1	y_1	
	2	+1	+1	-1	-1	+1	y_2	
	3	+1	-1	+1	-1	+1	y_3	
	4	+1	+1	+1	+1	+1	y_4	
Звезд-	5	+1	+1,414	0	0	2	0	y_5

ные точки	6	+1	-1,414	0	0	2	0	y ₆
	7	+1	0	+1,414	0	0	2	y ₇
	8	+1	0	-1,414	0	0	2	y ₈
Центр плана	9	+1	0	0	0	0	0	y ₉
	10	+1	0	0	0	0	0	y ₁₀
	11	+1	0	0	0	0	0	y ₁₁
	12	+1	0	0	0	0	0	y ₁₂
	13	+1	0	0	0	0	0	y ₁₃

Учитывая специфический характер ротатабельного плана в общем виде, можно также получить формулы для расчета коэффициентов уравнения регрессии и их дисперсий:

$$b_0 = \frac{A}{n} \left[2\lambda^2(k+2)(oy) - 2\lambda c \sum_{i=1}^k (i iy) \right]; \quad (I)$$

$$b_i = (c/n)(iy); \quad (II)$$

$$b_{ii} = \frac{A}{n} \left[c^2 [(k+2)\lambda - k](i iy) + c^2(1-\lambda) \sum_{i=1}^k (i iy) - 2\lambda c(oy) \right]; \quad (III)$$

$$b_{iu} = \frac{c^2}{n\lambda} (iu y); \quad (IV)$$

$$S_{b_0}^2 = \frac{2A\lambda^2(k+2)}{n} S_{\text{восп}}^2; \quad (V)$$

$$S_{b_{ii}}^2 = \frac{A[(k+1)\lambda - (k-1)c^2]}{n} S_{\text{восп}}^2; \quad (VI)$$

$$S_{b_{iu}}^2 = \frac{c^2}{\lambda n} S_{\text{восп}}^2; \quad (VII)$$

$$\text{где } (oy) = \sum_{j=1}^n x_{0j} y_j; \quad (iu y) = \sum_{j=1}^n x_{ij} x_{uj} y_j; \quad (VIII)$$

$$(iy) = \sum_{j=1}^n x_{ij} y_j; \quad (i iy) = \sum_{j=1}^n x_{ij}^2 y_j; \quad (IX)$$

$$C = \frac{n}{\sum_{j=1}^n x_{ij}^2}; \quad A = \frac{1}{2\lambda[(k+2)\lambda - k]}; \quad (X)$$

$$\lambda = \frac{nk}{(k+2)n_1} = \frac{k(n_1 + n_0)}{(k+2)n_1}. \quad (XI)$$

Здесь n_0 – число опытов в центре плана; n_1 – число остальных опытов.

Матрица ротатабельного планирования, оказывается неортогональной, так как

$$\sum_{j=1}^n x_{0j} \cdot x_{uj}^2 \neq 0; \quad \sum_{j=1}^n x_{ij}^2 \cdot x_{uj}^2 \neq 0; \quad i \neq u. \quad (6.33)$$

Следовательно, если какой-либо из квадратичных эффектов оказался незначимым, то после его исключения коэффициенты уравнения регрессии необходимо пересчитать заново.

При использовании ротатабельных планов второго порядка дисперсию воспроизводимости можно определить по опытам в центре плана. В связи с этим при проверке адекватности уравнения регрессии, полученного по ротатабельному плану второго порядка, поступают следующим образом.

Находят остаточную сумму квадратов

$$S_1^2 = \sum_{j=1}^n (y_j - \hat{y}_j)^2 \quad (6.34)$$

с числом степеней свободы

$$m_1 = n - l = n - \frac{(k+2)(k+1)}{2}.$$

По опытам в центре плана определяют сумму квадратов воспроизводимости

$$S_2^2 = \sum_{j=1}^{n_0} (y_{0j} - \bar{y}_{0j})^2 \quad (6.35)$$

с числом степеней свободы $m_2 = n_0 - 1$.

Далее находят сумму квадратов, характеризующих неадекватность

$S_3^2 = S_1^2 - S_2^2$, число степеней свободы которой

$$m_3 = m_1 - m_2 = n - \frac{(k+2)(k+1)}{2} - (n_0 - 1).$$

Проверяют адекватность по F-критерию:

$$F = \frac{S_3^2/m_3}{S_2^2/m_2}. \quad (6.36)$$

Уравнение адекватно, если $F < F_{\alpha; m_3; m_2}$.

Если модель второго порядка оказалась неадекватной, следует повторить эксперименты на меньшем интервале варьирования факторов или перенести центр плана в другую точку факторного пространства. В тех случаях, когда адекватность модели по-прежнему не достигается, рекомендуется перейти к планам третьего порядка.

6.4.3. Исследование причин образования расслоений в горячекатаных листах

В качестве примера планирования эксперимента второго порядка рассмотрим задачу исследования причин образования расслоений в горячекатаных листах (за основу числового материала взяты данные из книги¹).

Известно, что при прокатке листов толщиной более 12 мм появление брака связано большей частью с дефектами, унаследованными от слитка. Наиболее серьезными дефектами толстого листа являются расслоения, трещины и рванины.

Существует достаточно тесная связь между некоторыми параметрами выплавки стали и пораженностью листов расслоениями. По результатам ультразвуковой дефектоскопии было установлено, что на пораженность листов расслоениями (которая количественно может быть выражена в относительных единицах по площади расслоений, отнесенной к площади всего раската – Y , %) наиболее существенно влияют такие два фактора, как скорость выгорания углерода в период рудного кипения – x_1 , %/ч, и время разливки стали – x_2 , мин.

Уровни варьирования факторов представлены в табл. 6.13.

Таблица 6.13

¹ Паршин В.А., Зудов Е.Г., Колмогоров В.Л. Деформируемость и качество. – М.:Металлургия, 1979.

Уровни варьирования факторов

Уровни факторов	Факторы	
	x_1 , %/ч	x_2 , мин
Основной (нулевой)	0,35	5,5
Нижний	0,20	3,5
Верхний	0,50	7,5
Интервал варьирования	0,15	2,0

Для проведения эксперимента использовался ортогональный план второго порядка с тремя опытами в центре плана. По табл. 6.9 при числе факторов $k=2$ и $n_0 = 3$ величина звездного плеча составляет $\alpha = 1,148 \approx 1,15$, поэтому матрица планирования выглядит (табл.6.14) следующим образом:

Таблица 6.14

Ортогональный план второго порядка для двух факторов и с тремя опытами в центре плана

	Факторы (кодированные значения)		Факторы (натуральные значения)		Отклик y_j , %	
	x_1	x_2	x_1 , %/ч	x_2 , мин		
Ядро плана	1	-1	-1	0,20	3,5	0,36
	2	+1	-1	0,50	3,5	0,51
	3	-1	+1	0,20	7,5	1,33
	4	+1	+1	0,50	7,5	1,51
Звезд- ные точки	5	$\alpha = +1,15$	0	0,52	5,5	0,50
	6	$\alpha = -1,15$	0	0,18	5,5	0,31
	7	0	$\alpha = +1,15$	0,35	7,8	1,59
	8	0	$\alpha = -1,15$	0,35	3,2	0,45
Центр плана	9	0	0	0,35	5,5	0,30
	10	0	0	0,35	5,5	0,29
	11	0	0	0,35	5,5	0,31

Для обработки результатов эксперимента используем методику, изложенную ранее в параграфе 6.4.1.

Начнем с того, что к кодированным значениям x_1 и x_2 в исходную таблицу плана (см.табл.6.14) добавим фиктивный столбец $x_0=1$, а также дополнительные столбцы $x_{12}=x_1x_2$.

$$x'_{1j} = x_{1j}^2 - \frac{\sum_{j=1}^{11} x_{1j}^2}{11} \quad \text{и} \quad x'_{2j} = x_{2j}^2 - \frac{\sum_{j=1}^{11} x_{2j}^2}{11}.$$

Поскольку

$$\sum_{j=1}^{11} x_{1j}^2 = \sum_{j=1}^{11} x_{2j}^2 = (-1)^2 + (+1)^2 + (-1)^2 + (+1)^2 + (+1,15)^2 + (-1,15)^2 + (0)^2 + \dots + (0)^2 = 6,636,$$

$$\text{то } x'_{1j} = x_{1j}^2 - \frac{6,636}{11} \approx x_{1j}^2 - 0,6, \text{ аналогично } x'_{2j} \approx x_{2j}^2 - 0,6.$$

В итоге матрица плана эксперимента с фактическими результатами эксперимента y_j , полученными в ходе проведения опытов, выглядит следующим образом (табл.6.15):

Таблица 6.15

Матрица ортогонального плана второго порядка в кодированных значениях

Номер опыта	Ф а к т о р ы						Результат	
	x_0	x_1	x_2	x_{12}	x'_1	x'_2	y_j	\hat{y}_j
1	+1	-1	-1	+1	+0,4	+0,4	0,36	0,366
2	+1	+1	-1	-1	+0,4	+0,4	0,51	0,526
3	+1	-1	+1	-1	+0,4	+0,4	1,33	1,346
4	+1	+1	+1	+1	+0,4	+0,4	1,51	1,506
5	+1	+1,15	0	0	+0,7225	-0,6	0,50	0,507
6	+1	-1,15	0	0	+0,7225	-0,6	0,31	0,323
7	+1	0	+1,15	0	-0,6	+0,7225	1,59	1,587
8	+1	0	-1,15	0	-0,6	+0,7225	0,45	0,46
9	+1	0	0	0	-0,6	-0,6	0,30	0,296
10	+1	0	0	0	-0,6	-0,6	0,29	0,296
11	+1	0	0	0	-0,6	-0,6	0,31	0,296

В табл. 6.15 также приведены значения оценок отклика \hat{y}_j , найденные по модели, построенной после обработки экспериментальных данных.

Для оценки коэффициентов в уравнение регрессии

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_{12} + b'_{11} x'_1 + b'_{22} x'_2$$

по зависимостям (6.27) сформируем еще две дополнительные таблицы (табл. 6.16, 6.17), в которых представим результаты всех необходимых промежуточных расчетов.

Таблица 6.16

Таблица произведений кодированных значений факторов на значения отклика

Номер опыта	Произведения факторов на отклик					
	$x_0 \cdot y$	$x_1 \cdot y$	$x_2 \cdot y$	$x_{12} \cdot y$	$x_1' \cdot y$	$x_2' \cdot y$
1	0,36	-0,36	-0,36	0,36	0,144	0,144
2	0,51	0,51	-0,51	-0,51	0,204	0,204
3	1,33	-1,33	1,33	-1,33	0,532	0,532
4	1,51	1,51	1,51	1,51	0,604	0,604
5	0,5	0,575	0	0	0,36125	-0,3
6	0,31	-0,3565	0	0	0,223975	-0,186
7	1,59	0	1,8285	0	-0,954	1,148775
8	0,45	0	-0,5175	0	-0,27	0,325125
9	0,3	0	0	0	-0,18	-0,18
10	0,29	0	0	0	-0,174	-0,174
11	0,31	0	0	0	-0,186	-0,186
Σ -сумма	7,46	0,5485	3,281	0,03	0,305225	1,9319

Таблица 6.17

Величина квадратов кодированных значений факторов

Номер опыта	Квадраты факторов					
	$(x_0)^2$	$(x_1)^2$	$(x_2)^2$	$(x_{12})^2$	$(x_1')^2$	$(x_2')^2$
1	1	1	1	1	0,16	0,16
2	1	1	1	1	0,16	0,16
3	1	1	1	1	0,16	0,16
4	1	1	1	1	0,16	0,16
5	1	1,3225	0	0	0,522006	0,36
6	1	1,3225	0	0	0,522006	0,36
7	1	0	1,3225	0	0,36	0,522006
8	1	0	1,3225	0	0,36	0,522006
9	1	0	0	0	0,36	0,36
10	1	0	0	0	0,36	0,36
11	1	0	0	0	0,36	0,36
Σ -сумма	11	6,645	6,645	4	3,484013	3,484013

Используя значения из строк « Σ -сумма» в этих двух таблицах, находим оценки коэффициентов регрессии:

$$b_0' = 7,46/11 = 0,678182 ;$$

$$b_1 = 0,5485/6,645 = 0,082543 ;$$

$$b_2 = 3,281/6,645 = 0,493755 ;$$

$$b_{12} = 0,03/4 = 0,0075 ;$$

$$b_{11}' = 0,305225/3,484013 = 0,087607 ;$$

$$b_{22}' = 1,9319/3,484013 = 0,554504.$$

Значимость коэффициентов проверяем по критерию Стьюдента $t_i = |b_i|/S_{b_i}$, для чего находим дисперсию воспроизводимости $S_{\text{восп}}^2$ по трем параллельным опытам в центральной точке плана:

$$S_{\text{восп}}^2 = \frac{\sum_{i=1}^3 (y_{0i} - \bar{y}_0)^2}{3-1} = \frac{1}{3-1} \left[\sum_{i=1}^3 y_{0i}^2 - \frac{1}{3} \left(\sum_{i=1}^3 y_{0i} \right)^2 \right] =$$

$$= \frac{1}{2} \left[((0,30)^2 + (0,29)^2 + (0,31)^2) - \frac{1}{3} (0,30 + 0,29 + 0,31)^2 \right] = 0,0001$$

и рассчитываем по (6.28) дисперсии и средние квадратичные отклонения по каждому из коэффициентов:

$$S_{b_0}^2 = 0,0001/11 = 9,09091 \cdot 10^{-6}; \quad S_{b_0'} = \sqrt{9,09091 \cdot 10^{-6}} = 0,003015;$$

$$S_{b_1}^2 = S_{b_2}^2 = 0,0001/6,645 = 1,5 \cdot 10^{-5}; \quad S_{b_1} = S_{b_2} = \sqrt{1,5 \cdot 10^{-5}} = 0,003879;$$

$$S_{b_{12}}^2 = 0,0001/4 = 2,5 \cdot 10^{-5}; \quad S_{b_{12}} = \sqrt{2,5 \cdot 10^{-5}} = 0,005;$$

$$S_{b_{11}'}^2 = S_{b_{22}'}^2 = 0,0001/3,484013 = 2,87 \cdot 10^{-5}; \quad S_{b_{11}'} = S_{b_{22}'} = \sqrt{2,87 \cdot 10^{-5}} = 0,005357.$$

Тогда критерий Стьюдента по каждому из коэффициентов составит:

$$t_{b_0'} = 0,678182/0,003015 = 224,9;$$

$$t_{b_1} = 0,082543/0,003879 = 21,2;$$

$$t_{b_2} = 0,493755/0,003879 = 127,3;$$

$$t_{b_{12}} = 0,0075/0,005 = 1,5;$$

$$t_{b_{11}'} = 0,087607/0,005357 = 16,3;$$

$$t_{b_{22}'} = 0,554504/0,005357 = 103,5.$$

Поскольку критическое значение $t_{0,05;3-1}=4,30$ (**СТЮДРАСПОБР**(0,05;2)=4,302655725), то все коэффициенты в уравнении регрессии можно считать значимыми, кроме b_{12} .

Следовательно, окончательно уравнение регрессии можно записать в виде

$$\hat{y} = 0,68 + 0,08 \cdot x_1 + 0,49 \cdot x_2 + 0,09 \cdot x_1' + 0,55 \cdot x_2'.$$

Оценки отклика \hat{y}_j , полученные по этому уравнению для точек плана эксперимента, приведены в табл. 6.15. Сопоставляя полученные величины \hat{y}_j с опытными данными y_j , находим дисперсию адекватности, учитывая, что число значимых коэффициентов в уравнении регрессии равно пяти:

$$S_{\text{ад}}^2 = \frac{\sum_{j=1}^{11} (y_j - \hat{y}_j)^2}{11 - 5} = \frac{1}{6} \left[(0,360 - 0,366)^2 + (0,510 - 0,526)^2 + (1,330 - 1,346)^2 + \dots \right] = 0,00019.$$

Адекватность уравнения проверяется по критерию Фишера $F = S_{\text{ад}}^2 / S_{\text{восп}}^2 = 0,00019 / 0,0001 = 1,9$. Уравнение адекватно, поскольку составленное таким образом F -отношение меньше теоретического $F < F_{0,05; m_1=6; m_2=2} = 19,3$ (**ФРАСПОБР**(0,05;6;2) = 19,3294909), где $m_1 = 11 - 5 = 6$ – число степеней свободы дисперсии адекватности; $m_2 = 3 - 1 = 2$ – число степеней свободы дисперсии воспроизводимости.

В завершение данного примера перепишем полученное уравнение регрессии относительно x_1^2 и x_2^2 , учтя ранее введенные соотношения для $x_1' \approx x_1^2 - 0,6$ и $x_2' \approx x_2^2 - 0,6$.

$$\begin{aligned} \hat{y} &= 0,68 + 0,08 \cdot x_1 + 0,49 \cdot x_2 + 0,09 \cdot (x_1^2 - 0,6) + 0,55 \cdot (x_2^2 - 0,6) = \\ &= (0,68 - 0,09 \cdot 0,6 - 0,55 \cdot 0,6) + 0,08 \cdot x_1 + 0,49 \cdot x_2 + 0,09 \cdot x_1^2 + 0,55 \cdot x_2^2 = \\ &= 0,29 + 0,08 \cdot x_1 + 0,49 \cdot x_2 + 0,09 \cdot x_1^2 + 0,55 \cdot x_2^2. \end{aligned}$$

Проанализируем полученное уравнение на экстремум относительно x_1 и x_2 :

$$\begin{aligned} \frac{\partial \hat{y}}{\partial x_1} &= 0,08 + 2 \cdot 0,09 \cdot x_1 = 0; \\ \frac{\partial \hat{y}}{\partial x_2} &= 0,49 + 2 \cdot 0,55 \cdot x_2 = 0. \end{aligned}$$

Решая два последних соотношения, находим, что экстремум будет достигаться в точке с кодированными координатами $x_1 = -0,08/(2 \cdot 0,09) = -0,44$ и $x_2 = -0,55/(2 \cdot 0,49) = -0,44$, при этом натуральные значения факторов могут быть найдены из соотношений

$$\frac{X_1^{\text{нат}} - 0,35}{0,15} = X_1^{\text{код}} = -0,44;$$

$$\frac{X_2^{\text{нат}} - 5,5}{2} = X_2^{\text{код}} = -0,44.$$

Следовательно, при скорости выгорания углерода в период рудного кипения, равной $0,35 - 0,44 \cdot 0,15 = 0,28$ %/ч, и времени разлива стали $5,5 - 0,44 \cdot 2 = 4,6$ мин пораженность листов расслоениями будет минимальной, и составит порядка 0,16%.

$$\hat{y} = 0,29 + 0,08 \cdot (-0,44) + 0,49 \cdot (-0,44) + 0,09 \cdot (-0,44)^2 + 0,55 \cdot (-0,44)^2 \approx 0,16.$$

6.5. Планирование экспериментов при поиске оптимальных условий

Во многих случаях инженерной практики перед исследователем возникает задача не только выявления характера связи между двумя или несколькими рядами наблюдений, но и нахождения таких численных значений факторов, при которых отклик (выходной параметр) достигает своего экстремального значения (максимума или минимума). Эксперимент, решающий эту задачу, называется экстремальным. В этом случае задача сводится к оптимизационной и формулируется следующим образом: требуется определить такие координаты экстремальной точки $(x_1^*, x_2^*, \dots, x_k^*)$ поверхности отклика $y=f(x_1, x_2, \dots, x_k)$, в которой она максимальна (минимальна): $\max y(x_1, x_2, \dots, x_k) = y(x_1^*, x_2^*, \dots, x_k^*)$.

Графическая интерпретация задачи оптимизации объекта $y(x_1, x_2)$ при двух факторах x_1, x_2 представлена на рис. 6.4 а, б. Здесь точка А соответствует оптимальным значениям факторов x_1^* и x_2^* , обеспечивающим максимум функции отклика y_{\max} . Замкнутые линии на рис. 6.4, б характеризуют линии постоянного уровня и описываются уравнением $y=f(x_1, x_2)=B=\text{const}$.

Необходимость в экстремальных экспериментах довольно часто возникает в инженерной практике. Так, на модели шахтной печи с противоточно дви-

жущимся плотным продуваемым слоем, схема которой представлена на рис.6.5, требуется определить расположение фурмы по высоте печи H , ее диаметр D и высов L , обеспечивающие максимальную степень использования теплового потенциала газового потока. В данном случае факторами являются H , D , L , а в качестве функции отклика $y(H, D, L)$ в первом приближении можно использовать температуру отходящих из печи газов.

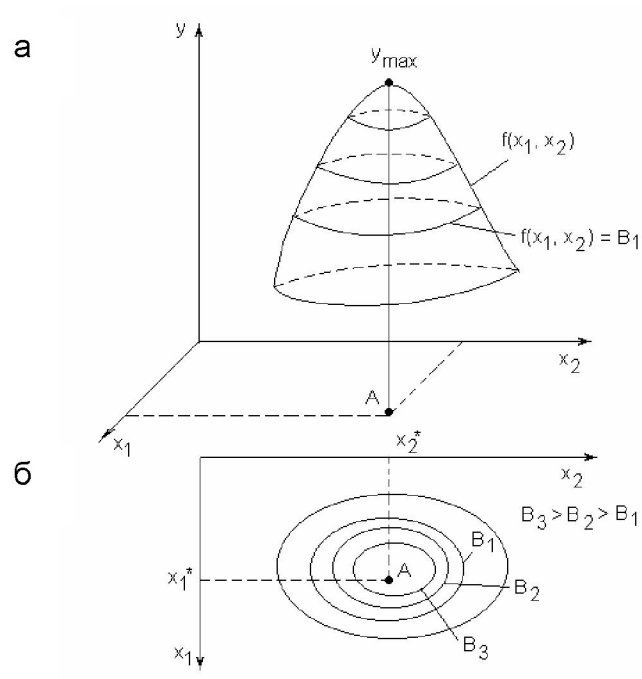


Рис.6.4. Поверхность отклика (а) и линии равного уровня (б): $y=f(x_1,x_2)=B=const$ для $n=2$

Заметим, что вид функции отклика в этом случае исследователю заранее неизвестен, т.е. отсутствует математическая модель, адекватно описывающая данный процесс. Требуется с наименьшими затратами (при минимальном числе опытов) определить оптимальные значения H^* , L^* , D^* , при которых температура отходящих газов минимальна.

Известный из практики метод "проб" и "ошибок", в котором факторы изменяются на основании опыта, интуиции или наугад, при обычно имеющем место значительном числе факторов при исследовании процессов в металлургии зачастую оказывается малоэффективным вследствие весьма сложной зависимости функции отклика от факторов.

Требуют значительно меньшего числа опытов и быстрее приводят к цели те поисковые методы оптимизации, где шаговое варьирование факторами производится целенаправленно по определенному плану. Поисковые методы оптимизации относятся к классу итерационных процедур, при этом весь процесс разбивается на шаги, на каждом шаге делается ряд опытов и определяется, каким образом нужно изменить факторы, влияющие на процесс, чтобы получить улучшение результата. При этом на каждом очередном шаге получаемая информация используется для выбора последующего шага.

Разработано множество методов пошаговой оптимизации, которые подробно рассматриваются в разделе вычислительной математики – "Численные методы оптимизации". Мы же рассмотрим только некоторые из них, эффективность использования которых в промышленном и лабораторном эксперименте применительно к металлургическим процессам подтверждена практикой.

6.5.1. Метод покоординатной оптимизации

Процесс поиска оптимума методом покоординатной оптимизации в графическом виде для двумерного случая представлен на рис.6.6. По этому методу выбирается произвольная точка M_0 и определяются ее координаты. Поиск оптимума осуществляется поочередным варьированием каждого их факторов. При этом сначала изменяют один фактор (x_1) при фиксированных остальных ($x_2 = \text{const}$) до тех пор, пока не прекращается прирост функции отклика (точка

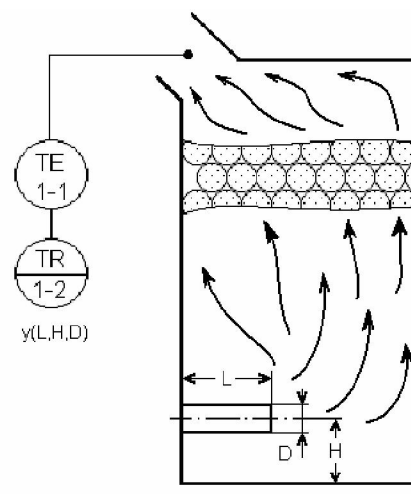


Рис. 6.5. Схема шахтной печи: 1-1 - датчик температуры; 1-2 – регистрирующий прибор

M_1). В дальнейшем изменяется другой фактор (x_2) при фиксированных остальных ($x_1 = \text{const}$), и далее процедура повторяется.

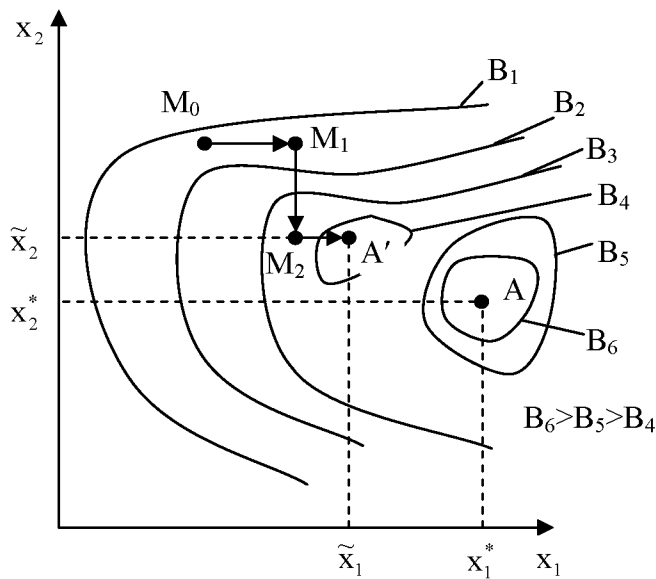


Рис.6.6. К методу покоординатной оптимизации

натных осей x_1 и x_2 вызывает уменьшение y . В результате решения находится ложный экстремум, находящийся в точке A' с координатами $\tilde{x}_1; \tilde{x}_2$, в то время как действительное значение максимума y_{\max} находится в точке A с координатами x_1^* и x_2^* .

В дальнейшем рассмотрим более совершенные методы.

Данный метод весьма прост, однако при большом числе факторов требуется значительное число опытов, чтобы достичь координат оптимума. Более того, при некоторых зависимостях $y=f(x_1, \dots, x_k)$ этот метод может привести к ложному результату. На рис.6.6 показан один из таких частных случаев, когда поочередное изменение каждого из факторов в любую сторону вдоль координатных осей вызывает уменьшение y .

6.5.2. Метод крутого восхождения

Известно, что кратчайший, наиболее короткий путь — это движение по градиенту, т.е. перпендикулярно линиям равного уровня, на которых функция отклика принимает постоянные значения $y(x_1, x_2, \dots, x_k) = V$. В связи с этим при оптимизации процесса рабочее движение целесообразно совершать в направлении наиболее быстрого возрастания функции отклика, т.е. в направлении градиента функции y .

Существуют различные модификации градиентного метода, одним из них является метод крутого восхождения. Сущность этого метода также рассмотрим на примере двухфакторной задачи (рис.6.7).

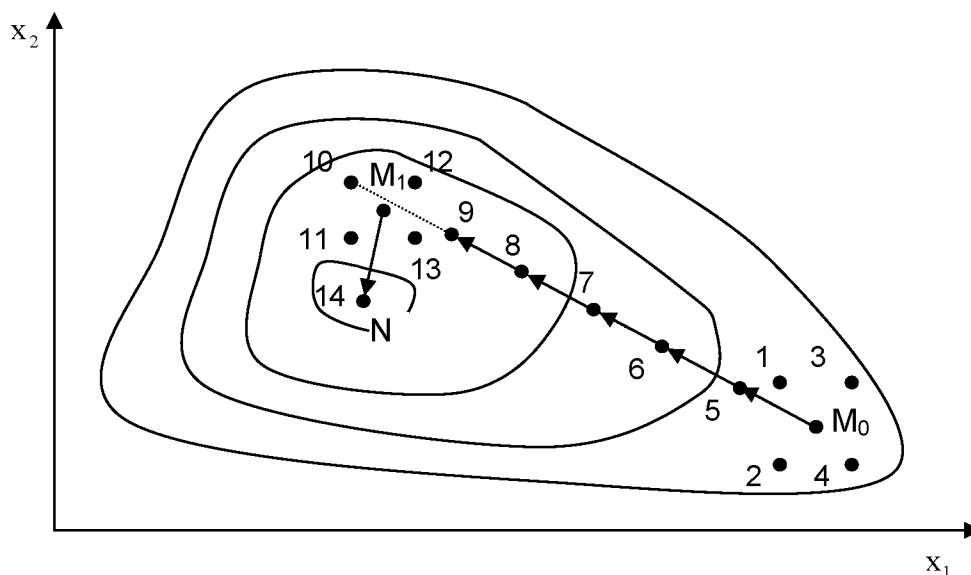


Рис. 6.7. Процедура оптимизации методом крутого восхождения

В этом случае шаговое движение осуществляется в направлении наискорейшего возрастания функции отклика, т.е. $\text{grad } y(x_1, x_2)$. Однако направление корректируют не после каждого следующего шага, а при достижении в некоторой точке на данном направлении частного экстремума функции отклика.

Пусть в окрестности точки M_0 как центра плана поставлен ПФЭ 2^2 . Координаты отдельных опытов соответствуют точкам 1-4. По результатам ПФЭ можно рассчитать коэффициенты линейного уравнения регрессии.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2.$$

Градиент функции отклика в этой точке определяется как

$$\text{grad } y = \frac{\partial y}{\partial x_1} \cdot \vec{i} + \frac{\partial y}{\partial x_2} \cdot \vec{j}, \quad (6.37)$$

где \vec{i}, \vec{j} — единичные векторы в направлении координатных осей.

Следовательно, для движения по градиенту необходимо изменять факторы пропорционально их коэффициентам регрессии и в сторону, соответствующую знаку коэффициента. В процессе поиска двигаются в этом направле-

нии до тех пор, пока не будет обнаружен локальный максимум (точка M_1 на рис. 6.7). В точке последнего находят новое направление градиента (направление M_1N), осуществляя опять же ПФЭ, и далее процедура повторяется. Стрелками на рис. 6.7 показана траектория движения к оптимуму.

Практически алгоритм сводится к следующей последовательности операций.

1. Планирование и постановка ПФЭ (или ДФЭ) в окрестности точки начального состояния x_{i0} . Расчет коэффициентов b_i линейной математической модели с целью определения направления градиента.

2. Расчет произведений $b_i \Delta x_i$, где Δx_i — интервалы варьирования факторов при ПФЭ (ДФЭ).

3. Выбор базового фактора $x_i = x_{i0}$, у которого $|b_i \Delta x_i| = a = \max$.

4. Выбор шага крутого восхождения для базового фактора h_a .

Этот выбор производится на основании имеющейся априорной информации или с учетом опыта исследователя, технологических соображений или других критериев. Относительно выбора шага заметим, что слишком малый шаг потребует значительного числа опытов при движении к оптимуму, а большой шаг создает опасность проскакивания области оптимума.

5. Расчет шагов изменения других факторов по формуле

$$h_i = (b_i \Delta x_i) h_a / a. \quad (6.38)$$

Это соотношение между величинами шагов изменения отдельных факторов обеспечивает движение по градиенту в факторном пространстве.

6. Составление плана движения по градиенту. Для этого в соответствии с определенными значениями шагов изменения факторов и их последовательным алгебраическим суммированием с основным уровнем в точке

$$x_{ik} = x_{i0} + kh_i, \quad k = 1, 2, \dots$$

находят координаты опытов 5, 6, 7, 8, 9, 10 (см. рис. 6.7). Часть этих опытов полагают "мысленными". "Мысленный" опыт заключается в получении предсказанных (расчетных) значений функции отклика по линейному уравнению рег-

рессии, что позволяет сократить объем реальных опытов, т.е. увеличить скорость продвижения к экстремуму. При "мысленном эксперименте" перевод координат в кодированную форму и подстановка их в уравнение модели объекта должна подтвердить действительное возрастание y . Обычно реальные опыты в начале движения из базовой точки вдоль направления градиента ставятся через 2-4 мысленных опыта. Другие опыты реализуют на практике, определяя последовательность значений y в направлении градиента. Из опытных данных находят положение локального экстремума (точка M_1 на рис.6.7).

7. В окрестности локального экстремума ставят новую серию опытов (ПФЭ илиДФЭ) для определения новых значений коэффициентов уравнения регрессии и нового направления градиента (направление M_1N на рис.6.7). В дальнейшем процедура повторяется до достижения следующего локального экстремума и так далее вплоть до определения окрестности координат максимума функции отклика, которая носит название почти стационарной области.

Признаком достижения этой области является статистическая незначимость коэффициентов b_i . В почти стационарной области становятся значимы эффекты взаимодействия и квадратичные эффекты. Здесь требуется переходить отДФЭ (если он использовался ранее) к ПФЭ, а если и этого окажется недостаточно, перейти от планов эксперимента первого порядка к планам второго порядка.

Очевидно, что в задачах, где требуется определить координаты не максимума, а минимума функции отклика, знаки коэффициентов b_i следует поменять на обратные. В этом случае движение в факторном пространстве осуществляется по направлению, противоположному вектору градиента.

6.5.3. Симплексный метод планирования

Метод симплексного планирования позволяет без предварительного изучения влияния факторов найти область оптимума. В этом методе не требуется вычисления градиента функции отклика, поэтому он относится к безградиентным методам поиска оптимума. Для этого используется специальный план эксперимента в виде симплекса.

Симплекс — это простейший выпуклый многогранник, образованный $k+1$ вершинами в k -мерном пространстве, которые соединены между собой прямыми линиями. При этом координаты вершин симплекса являются значениями факторов в отдельных опытах. Так, например, в двухфакторном пространстве (на плоскости) $k=2$ симплекс — любой треугольник, в трехфакторном (трехмерном) $k=3$ пространстве — тетраэдр и т.д.

Симплекс называется правильным или регулярным, если все расстояния между образующими его вершинами равны (равносторонний треугольник, правильный тетраэдр и др.).

После построения исходного симплекса и проведения опытов при значениях факторов, соответствующих координатам его вершин, анализируют результаты и выбирают вершину симплекса, в которой получено наименьшее (наихудшее) значение функции отклика. Для движения к оптимуму необходимо поставить опыт в новой точке, являющейся зеркальным отображением точки с наихудшим (минимальным) результатом относительно противоположной грани симплекса. На рис.6.8 представлено геометрическое изображение симплекс-метода для двумерного случая $k=2$.

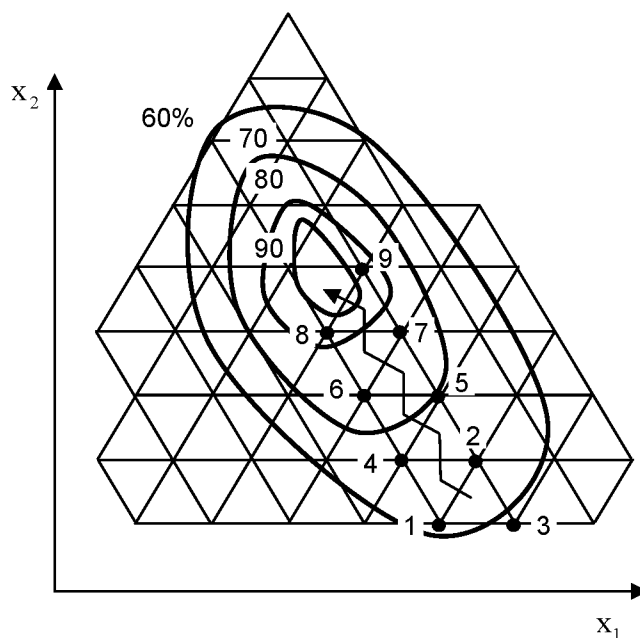


Рис. 6.8. Схема движения к оптимальной области симплексным методом

По итогам проведения опытов 1, 2 и 3 худшим оказался опыт 3. Следующий опыт ставится в точке 4, которая образует с точками 1 и 2 новый правиль-

ный симплекс. Далее сопоставляются результаты опытов 1, 2 и 4. Наихудший результат получен в точке 1, поэтому она в новом симплексе заменяется зеркальным отображением (точкой 5) и т.д., пока не будет достигнута почти стационарная область. Следует заметить, что хотя этот путь и зигзагообразен, общее число опытов, необходимых для достижения области оптимума, может быть небольшим за счет того, что проводить $k+1$ опыт приходится лишь в начале, а в дальнейшем каждый шаг сопровождается проведением только одного дополнительного опыта, условия которого выбираются на основе предшествующих результатов.

После изложения основных идей симплексного метода планирования оптимальных экспериментов остановимся на некоторых его деталях. Выбор размеров симплекса и его начального положения в известной степени произволен. Для построения начального симплекса значения в каждом опыте исходного симплекса определяются по формуле

$$x_{ij} = x_{i0} + C_{ij}\Delta x_i, \quad (6.39)$$

где x_{i0} — координаты центра начального симплекса; Δx_i — интервал варьирования i -го фактора; C_{ij} — кодированное значение i -го фактора для j -го опыта, выбираемые из числовой матрицы для симплексного планирования, приведенные в табл. 6.18.

Таблица 6.18

*Коэффициенты C_{ij} для выбора координат симплекса **

Номер опыта ($\downarrow j$)	Факторы ($\rightarrow i$)					
	x_1	x_2	x_3	...	x_{k-1}	x_k
1	k_1	k_2	k_3	...	K_{k-1}	K_k
2	$-R_1$...		K_k
3	0	$-R_2$...		K_k
4	0	0	$-R_3$...		K_k
...	K_k
k	0	0	0	0	R_{k-1}	K_k
$k+1$	0	0	0	0	0	R_k

$$*) k_i = \frac{1}{i+1} \sqrt{\frac{i+1}{2i}} = \sqrt{\frac{1}{2i(i+1)}}; R_i = \sqrt{\frac{i}{2(i+1)}}; i = 1, 2, \dots, k,$$

где k — число факторов

Если, например, необходимо составить симплекс-план для двух факторов, то вначале ставят три опыта со следующими координатами:

1-й опыт

$$x_{11} = x_{10} + k_1 \Delta x_1;$$

$$x_{21} = x_{20} + k_2 \Delta x_2.$$

2-й опыт

$$x_{12} = x_{10} - R_1 \Delta x_1;$$

$$x_{22} = x_{20} + k_2 \Delta x_2.$$

3-й опыт

$$x_{13} = x_{10} + 0;$$

$$x_{23} = x_{20} - R_2 \Delta x_2.$$

Симплекс, рассчитанный по этим формулам, представлен на рис.6.9.

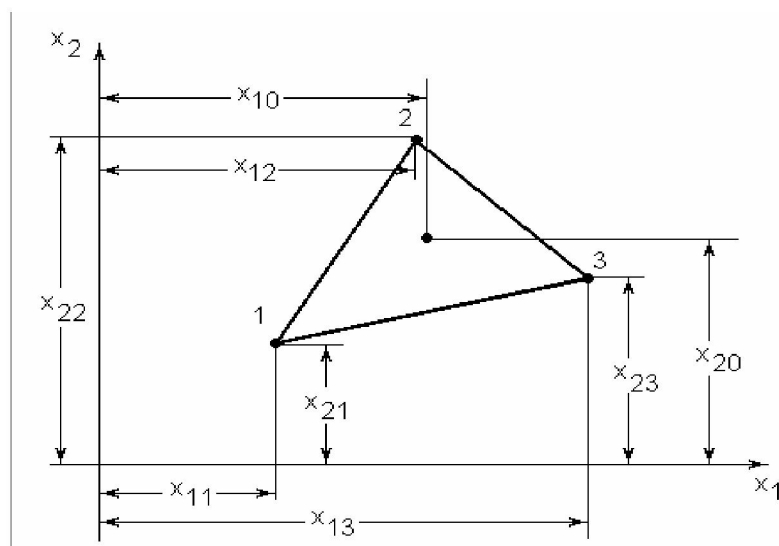


Рис. 6.9. Схема построения начального симплекса

Так, если $x_{10}=0$ и $x_{20}=0$, а $\Delta x_1=\Delta x_2=1$, то координаты опытов будут равны (см. рис.6.10): опыт 1 (0,5;0,289), опыт 2 (-0,5; 0,289) и опыт 3 (0;-0,577), что соответствует координатам вершин равностороннего треугольника с длиной стороны, равной 1. Начало координат в этом случае находится в точке пересечения медиан (биссектрис).

Для определения условий проведения опыта в отраженной точке (координат новой вершины симплекса) используется формула

$$x_{in} = \frac{2}{k} \sum_{j=1}^{k+1} x_{ij} - x_{i3}, \quad j \neq i_3, \quad (6.40)$$

где x_{in} — координата новой точки (новой вершины) симплекса для i -й переменной; x_{i3} — координата заменяемой точки (координата вершины симплекса с наихудшим откликом перед ее отбрасыванием); $\frac{1}{k} \sum_{j=1}^{k+1} x_{ij}$ — среднее значение из координат всех вершин симплекса, кроме заменяемой.

Известны следующие критерии окончания процесса последовательного отражения наихудших вершин и постановки очередных опытов в новых вершинах:

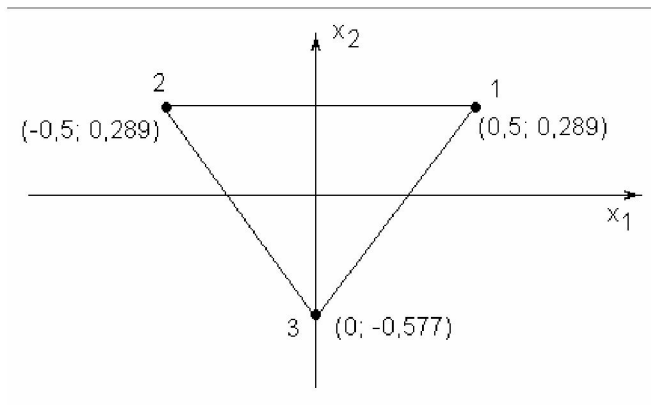


Рис.6.10. Координаты вершин симплекса при $x_{i0}=0$, $\Delta x_i=1$ и $n=2$

1. Разность значений функции отклика в вершинах симплекса становится меньше ранее заданной величины. Это означает либо выход в "почти стационарную" область вблизи оптимума, либо достижение участка поверхности $\hat{y} = f(x_1; \dots; x_k) = const$ в виде "плато". В этом случае дополни-

тельными опытами в стороне от симплекса следует удостовериться в отсутствии других участков с более существенной кривизной поверхности $y = f(x_1; \dots; x_k)$ и принять величину с экстремальным значением функции отклика за точку оптимума.

2. Отражение любой из вершин симплекса после однократного качания приводит к его возврату в прежнее положение. При этом есть основания утверждать "накрытие" симплексом точки оптимума.

3. Циклическое движение симплекса вокруг одной из его вершин на протяжении более чем нескольких шагов. Подобная ситуация имеет место, когда искомый оптимум располагается внутри области, охватываемой циркулирующим симплексом.

В случаях 2 и 3 рекомендуется уменьшить размеры симплекса, т.е. расстояния между вершинами, и продолжить поиск до желаемого уточнения координат искомого оптимума.

Изложенный алгоритм симплексного метода сравнительно прост, он достаточно эффективен, однако работает недостаточно быстро. Существует его модификация, известная под названием "метод деформируемого симплекса", которая ускоряет процесс поиска оптимума за счет использования на данном шаге информации, накопленной на предыдущих шагах.

Сущность метода поиска по деформированному симплексу заключается в том, что при отражении наихудшей вершины относительно центра тяжести противоположной грани размер симплекса не остается постоянным, а осуществляется его деформация (растяжение или сжатие). Для пояснения существа метода введем координату центра тяжести \bar{x}_i остальных (за исключением наихудшей) вершин симплекса:

$$\bar{x}_i = \sum_{j=1}^{k+1} x_{ij} / k; \quad j \neq i_3.$$

Тогда формула (6.40) может быть преобразована к виду

$$x_{iH} = 2\bar{x}_i - x_{i3}$$

или

$$\tilde{x}_{iH} = \bar{x}_i + \alpha(\bar{x}_i - x_{i3}).$$

При $\alpha=1$ получим выражение (6.40) и $\tilde{x}_{iH} = x_{iH}$.

Введем обозначения:

y_3 — наихудший (минимальный) отклик в симплексе;

y_{\max} — наилучший (максимальный) отклик;

y_3' — отклик, следующий за наихудшим.

Следовательно $y_3 < y_{\max} < y_3'$.

В зависимости от значения функции отклика в точке нормального отражения y_H при $\alpha=1$ возможны следующие варианты:

1) если $y_3 < y_n < y_{\max}$, т.е. x_{in} будет нехудшей и нелучшей точкой в новом наборе точек, то x_{i3} следует заменить на x_{in} с $\alpha=1$. В этом случае осуществляется нормальное отражение;

2) если $y_n > y_{\max}$, то x_{in} оказывается новой лучшей точкой в новом наборе точек. В этом случае направление растяжения признается “весьма удачным” и симплекс растягивается в нормальном направлении. Для этого случая $1 < \alpha < 2$ и α называется коэффициентом растяжения;

3) если $y_3 < y_n < y_3'$, то направление отражения признается правильным, но симплекс слишком велик и его следует сжать выбором коэффициента сжатия α из диапазона $0 < \alpha < 1$;

4) если $y_n < y_3$, то даже направление отражения выбрано неверно и следует осуществить отрицательное сжатие выбором отрицательного значения коэффициента α ($-1 < \alpha < 0$).

Таким образом, на каждом шаге следует вначале нормально отразить наихудшую вершину симплекса ($\alpha=1$), поставить в этой точке опыт, определить y_n , а затем поставить следующий опыт в точке факторного пространства \tilde{x}_n , координаты которой определяются по формуле (6.40) с учетом рассмотренных вариантов 1-4.

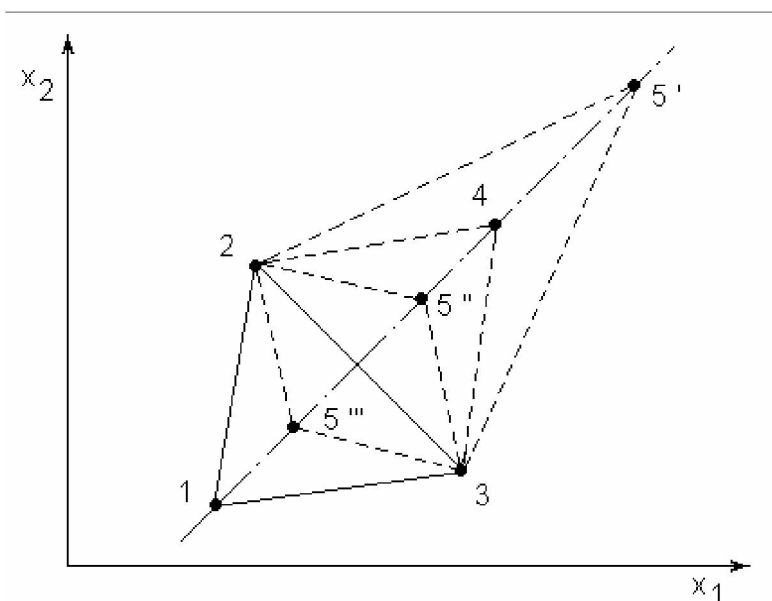


Рис.6.11. К методу деформированного симплекса

На рис.6.11 показаны точка 4 очередного опыта при нормальном отражении ($\alpha=1$) наихудшей вершины 1, точки 5', 5'', 5''' последующих опытов для случаев соответственно растяжения ($\alpha=2$), сжатия ($\alpha=0,5$) и отрицательного сжатия ($\alpha=-0,5$) симплекса.

Таким образом, метод поиска по деформированному симплексу обладает повышенной гибкостью, позволяющей учитывать особенности поверхности отклика.

Пример 6.1 [18]. Пусть объект обладает свойствами, соответствующими уравнению

$$y = 4 + 12x_1 - x_1^2 + 30x_2 - 3x_2^2.$$

Найдем экстремум функции симплекс-методом. Выберем основной уровень факторов. Предположим, что по некоторым данным экстремум находится вблизи значений $x_{10}=3$ и $x_{20}=-1$, которые и принимаем за основной уровень. Интервал варьирования примем равным $\Delta x_1=1$ и $\Delta x_2=1,5$. Найдем

$$k_1 = \sqrt{\frac{1}{2 \cdot 1 \cdot (1+1)}} = 0,5; \quad R_1 = \sqrt{\frac{1}{2 \cdot (1+1)}} = 0,5;$$

$$k_2 = \sqrt{\frac{1}{2 \cdot 2 \cdot (2+1)}} = 0,289; \quad R_2 = \sqrt{\frac{2}{2 \cdot (2+1)}} = 0,577.$$

Находим координаты первых трех опытов, так как $m+1=2+1=3$.

Вершина 1: $x_{11} = 3+0,5 \cdot 1=3,5$; $x_{21} = -1+0,289 \cdot 1,5= -0,565$;

Вершина 2: $x_{12} = 3-0,5 \cdot 1=2,5$; $x_{22} = -1+0,289 \cdot 1,5= -0,565$;

Вершина 3: $x_{13} = 3+0=3$; $x_{23} = -1-0,577 \cdot 1,5= -1,865$.

Подставляя найденные координаты вершин в уравнение, получили следующие результаты опыта: $y_1=15,84$; $y_2=9,78$; $y_3= -35,5$. Самый худший результат $y_3= -35,5$. Следовательно, условия опыта 3 следует заменить. Геометрическая траектория движения показана на рис. 6.12.

Вычисляем координаты вершины 4.

$$x_{14} = \frac{2(3,5+2,5)}{2} - 3 = 3; \quad x_{24} = \frac{2(-0,565-0,565)}{2} + 1,865 = 0,735. \quad \text{Результат} - y_4=52,1.$$

Сравнивая результаты y_1 , y_2 и y_4 , видим, что худший результат y_2 . Вычисляем координаты вершины 5:

$$x_{15} = 4; \quad x_{25} = 0,735; \quad y_5 = 57,1.$$

Вычисляем координаты вершины 6, заменяя вершину 1:

$$x_{16} = 3,5; \quad x_{26} = 2,035; \quad y_6 = 82,6.$$

Далее получим вершины 7 с координатами (4,5;2,035); 8 (4,3,3); 9 (5;3,3); 10 (4,5;4,8); 11 (5,5;4,8) и результаты последних трех опытов:

$$y_9 = 105; \quad y_{10} = 113; \quad y_{11} = 112,32.$$

Находим координаты вершины 12:

$$x_{112} = \frac{2(4,5 + 5,5)}{2} - 5 = 5; \quad x_{212} = \frac{2(4,6 + 4,6)}{2} - 3,3 = 5,9; \quad y_{12} = 111.$$

Как видно, координаты вершины 12 соответствуют худшим результатам, чем 10 и 11. Поэтому возвращаемся к предыдущему симплексу с вершинами 9,10, 11 и выбираем худший результат, не обращая внимания на опыт 9. Следовательно, заменить необходимо вершину 10:

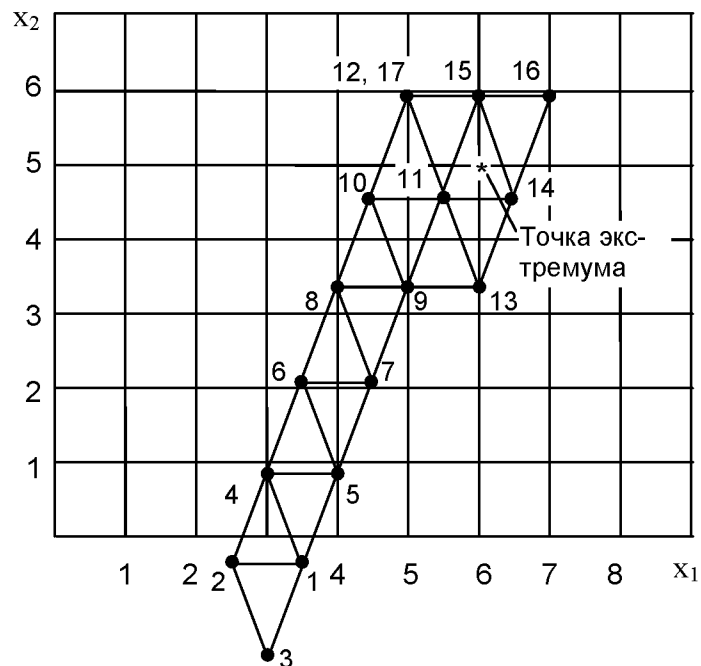


Рис. 6.12. К решению примера 6.1

$$x_{113} = \frac{2(5 + 5,5)}{2} - 4,5 = 6; \quad x_{213} = \frac{2(5,3 - 4,6)}{2} - 4,6 = 3,3; \quad y_{13} = 106.$$

В новом симплексе 9 (5;3,3); 10 (5,5;4,6) и 13 (6;3,3) худший результат у опыта 9. Заменяем вершину 9:

$$x_{114} = \frac{2(5,5 + 6)}{2} - 5 = 6,5; \quad x_{214} = \frac{2(4,6 + 3,3)}{2} - 3,3 = 4,6; \quad y_{14} = 114,21.$$

Результаты опытов 11, 13 и 14 следующие: $y_{11} = 112,32$, $y_{13} = 106$ и $y_{14} = 114,21$. Заменяем вершину 13:

$$x_{115} = \frac{2(5,5 + 6,5)}{2} - 6 = 6; \quad x_{215} = \frac{2(4,6 + 4,6)}{2} - 3,3 = 5,9; \quad y_{15} = 112.$$

Получен худший результат, чем в опытах 11 и 14. Поэтому заменяем опыт 11:

$$x_{116} = \frac{2(6,5 + 6)}{2} - 5,5 = 7; \quad x_{216} = \frac{2(4,6 + 5,9)}{2} - 4,6 = 5,9; \quad y_{16} = 111.$$

Это также худший результат, поэтому в симплексе 11, 14 и 15 заменяем опыт 14:

$$x_{117} = \frac{2(5,5 + 6)}{2} - 6,5 = 5; \quad x_{217} = \frac{2(4,6 + 5,9)}{2} - 4,6 = 5,9.$$

Вершины 17 и 12 совпадают, $y_{17} = y_{12} = 111$. Получен снова худший результат. Следовательно, экстремум находится внутри этого симплекса (см. рис. 6.12). Далее можно уменьшить интервал варьирования и от любой вершины двигаться вновь. Если же с точностью до шага варьирования результаты устраивают, можно считать задачу решенной.

Следовательно, координаты экстремума

$$x_1 \approx 6,5 \text{ и } x_2 \approx 4,6; \quad y = 114,21.$$

Истинные координаты экстремума

$$x_1 = 6 \text{ и } x_2 = 5; \quad y = 115.$$

7. КОМПЬЮТЕРНЫЕ МЕТОДЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ РЕЗУЛЬТАТОВ ИНЖЕНЕРНОГО ЭКСПЕРИМЕНТА

7.1. Общие замечания

В данной главе мы рассмотрим возможности использования отдельных компьютерных программ (пакетов прикладных программ, программных сред, компьютерных систем) для статистической обработки данных, полученных в ходе инженерного эксперимента. Преимущества использования в этой области компьютерных программных продуктов очевидны, однако сделаем некоторые замечания.

В настоящее время темпы развития компьютерных технологий настолько велики, что создаваемые аппаратные и программные средства обработки информации, в том числе и статистической, совершенствуются практически с каждым месяцем, приобретая все новые и новые возможности. С распространением мощных персональных компьютеров стало возможно реализовывать методы расчета, которые раньше считались очень трудоемкими в вычислениях. На рынке программного обеспечения существуют достаточно сложные пакеты прикладных программ, профессионально ориентированные на обработку статистической информации и позволяющие выявлять закономерности на фоне случайностей, делать обоснованные выводы и прогнозы, оценивать вероятности их выполнения. Эти программные среды обладают высокой степенью универсальности, а их применимость и технология использования практически не зависят от предметной области (металлургия, экономика, медицина и др.).

Тенденцией развития современных компьютерных технологий является объединение (интеграция) функций отдельных пакетов программ (математических, статистических, текстовых, графических, коммуникационных и др.) в так называемые интегрированные компьютерные среды. Эта особенность наиболее четко прослеживается с выходом новых версий популярных программных продуктов, когда возможности существующих программ расширяются за счет включения в них новых функций. В качестве примера можно привести пакет Microsoft Office, включающий в себя наряду со средствами создания и обработки текста (Word), баз данных (Access), презентаций (Power Point) также табличный процессор Excel, предназначенный, вообще говоря, для создания электронных таблиц и манипулирования их данными. В состав Microsoft Excel входит набор средств анализа данных (пакет анализа), предназначенный для решения сложных статистических задач. Для проведения анализа

данных с помощью этих средств достаточно указать (отметить) диапазон входных данных из таблицы и выбрать необходимые параметры; расчет будет проведен с помощью подходящей статистической функции, а результат будет помещен в выходной диапазон таблицы. Кроме того, специальные средства позволяют представить результаты в графическом виде. Для успешного применения процедур анализа в Microsoft Excel также необходимы соответствующие знания в области статистических расчетов, для которой эти инструменты были разработаны. Несмотря на то, что электронные таблицы уступают по своим возможностям специализированным пакетам статистической обработки данных, изучение возможностей и владение навыками работы с Microsoft Excel делает их мощным инструментом в руках инженера-исследователя.

Компьютерные системы для анализа данных – статистические пакеты (СП) – являются, по сравнению с другими наукоемкими программами, пожалуй, наиболее широко применяемыми в инженерной практике и исследовательской работе в разнообразных областях человеческой деятельности. Статистический пакет должен удовлетворять определенным требованиям, на которые в первую очередь надо обращать внимание при его выборе:

- использование простого пользовательского интерфейса, основанного на проблемно-ориентированном языке высокого уровня для формулировки задания пользователя;
- модульность программного обеспечения, автоматическая организация процесса обработки данных и связей между модулями пакета;
- развитая система поддержки при выборе способов обработки данных, визуальном отображении результатов и их интерпретации;
- наличие средств сохранения результатов проделанного анализа в виде графиков и таблиц;
- совместимость с другим программным обеспечением.

Современная программа анализа данных, в большинстве случаев, представляет собой электронные таблицы с ограниченными по сравнению с обычными электронными таблицами средствами манипулирования данными, но с достаточно мощными методами расчетов по этим данным. Общая технология статистического анализа данных с использованием статистического пакета включает в себя следующие основные этапы:

- 1) ввод данных в электронную таблицу с исходными данными и их предварительное преобразование перед анализом (структурирование, построение необходимых выборок, ранжирование и т. д.);
- 2) визуализация данных при помощи того или иного типа графиков;
- 3) определение подходящих методов статистической обработки;
- 4) применение конкретной процедуры статистической обработки;
- 5) вывод результатов анализа в виде графиков и электронных таблиц с числовой и текстовой информацией;
- 6) подготовка, печать и сохранение отчета.

Для расчетного анализа данных в СП используются отдельные библиотеки модулей. Модуль СП – это внешняя процедура или программа на языке программирования высокого уровня, удовлетворяющая некоторым дополнительным ограничениям, наиболее важными из которых являются: ограничения на способ аварийного завершения работы модуля; на способы связи по информации, например на допустимость переменных внешнего типа и использование общей области памяти; на возможность передачи управления между модулями с помощью операторов вызова, расположенных в теле модуля; на использование операторов ввода-вывода. Отметим наиболее типовые расчетные модули современных статических пакетов, которые условно разделим на следующие три группы:

- описательная статистика и разведочный анализ исходных данных;
- статистическое исследование зависимостей;
- вспомогательные программы.

Модуль описательной статистики и разведочного анализа исходных данных позволяет проводить:

- анализ резко выделяющихся наблюдений;
- проверку статистической независимости рядов наблюдений;
- определение основных числовых характеристик и частотную обработку исходных данных (построение гистограмм, полигонов частот, вычисление выборочных средних, дисперсий и т.д.);
- расчет критериев однородности (средних, дисперсий, законов распределения и т.д.);
- определение критериев согласия (хи-квадрат, Колмогорова–Смирнова и др.);
- статистическое оценивание параметров;

- вычисление наиболее распространенных законов распределения вероятностей (нормального, Пуассона, хи-квадрат и некоторых других);
- визуализацию анализируемых многомерных статистических данных.

Модуль статистического исследования зависимостей является достаточно объемной частью любого СП. Он включает в себя решение следующих задач:

- корреляционно-регрессионный анализ;
- дисперсионный анализ;
- планирование регрессионных экспериментов и выборочных обследований и др.

Вспомогательные программы расширяют возможности статистических пакетов и реализуют, в частности, оптимизационные алгоритмы, вычислительные процедуры, основанные на нейросетях и генетических алгоритмах, задачи статистического моделирования на ЭВМ, которые являются полезными составными элементами компьютерных имитационных экспериментов, используемых при анализе сложных реальных систем.

В настоящее время существует множество источников информации по использованию статистических пакетов как в виде книгопечатных изданий, так и в электронном виде, которые размещены на сайтах в сети Internet. Для более детального знакомства с процедурами компьютерной обработки результатов статистических исследований заинтересованным читателям можно рекомендовать специальный справочник [20]. Ниже в табл. 7.1 представлены адреса ресурсов Internet, на которых расположена информация по некоторым распространенным статистическим пакетам.

Таблица 7.1

Статистические пакеты

Адрес	Название программы	Разработчик
1	2	3
www.statsoft.ru	STATISTICA	StatSoft Inc., США
www.spss.ru	SPSS	SPSS Inc., США
www.stat-graphics.com	STATGRAPHICS Plus	Manugistics Inc.,
www.sas.com	StatView	SAS Institute Inc.,
www.ncss.com	NCSS	NCSS Statistical Software,
www.minitab.com	Minitab	Minitab Inc.

1	2	3
www.insightful.com	S-PLUS	Insightful Corp.
statsoft.msu.ru	STADIA	НПО «Информатика и компьютеры», Россия
www.softsite.ru	Stat-Media	Центр интеллектуальных систем «Метод», Россия
www.megaputer.ru	PolyAnalyst	«Мегапьютер Интеллидженс», Россия

Отметим, что многие поставщики предлагают пользователям пробные и демонстрационные версии статистических программ, как правило, отличающиеся тем, что пробная версия представляет собой полнофункциональный продукт с ограниченным сроком использования, а демонстрационная версия в большинстве случаев напоминает электронную презентацию.

Следующий параграф посвящен краткому описанию основных статистических функций электронных таблиц Microsoft Excel с комментариями по их использованию в теории инженерного эксперимента для статистической обработки экспериментальных данных и анализа результатов наблюдений. Некоторые из этих функций использовались нами в предыдущих главах данного пособия при изложении соответствующих разделов теории инженерного эксперимента и иллюстрации примеров. Предполагается, что читатель уже имеет некоторые навыки работы на компьютере в среде электронных таблиц Microsoft Excel, поэтому может использовать данный материал в качестве справочного пособия для статистического анализа данных.

Затем мы рассмотрим материал, посвященный знакомству с возможностями интегрированной системы статистического анализа и обработки данных STATISTICA. Наш выбор обусловлен тем фактом, что данная система является на сегодня одной из лидирующих на рынке программного обеспечения для статистической обработки данных. Кроме этого, важным моментом является наличие русскоязычной версии программы и множества публикаций, посвященных этому пакету, например [21].

7.2. Статистические функции Microsoft Excel

Пакет Microsoft Excel не предназначен для комплексного статистического анализа данных в отличие от специализированного статистического программного обеспечения, например, пакета STATISTICA. Однако и на базе электронных таблиц можно провести некоторую статистическую обработку данных для большинства ин-

женерных задач. Функции, реализующие статистические методы обработки и анализа данных, в Microsoft Excel реализованы в виде специального программного расширения – надстройки «Пакет анализа», которая входит в поставку данного программного продукта и может устанавливаться по желанию пользователя.

Установка надстройки «Пакет анализа» производится из меню «Сервис/Надстройки», после чего в диалоговом окне «Надстройки» (рис. 7.1) необходимо отметить флажок пункта «Пакет анализа» и нажать кнопку ОК.

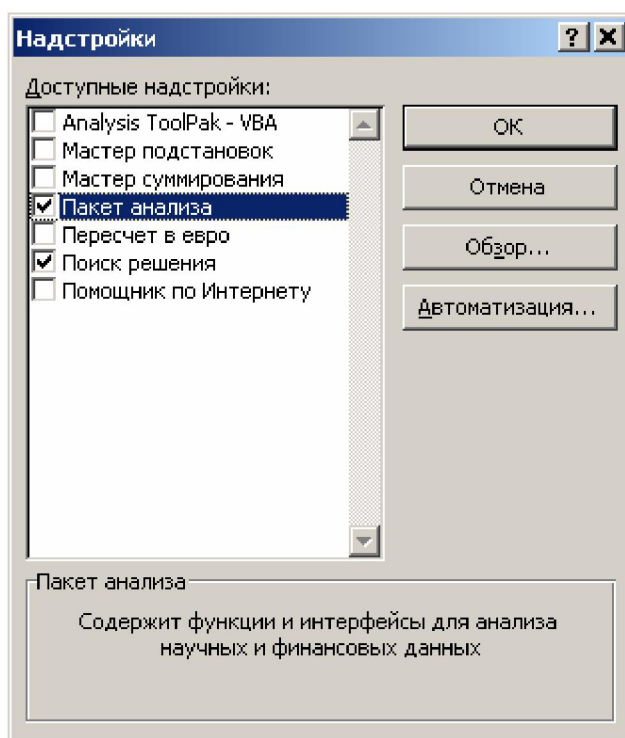


Рис. 7.1. Установка пакета анализа в Microsoft Excel

Если процесс установки завершен успешно, то в меню «Сервис» появляется еще один пункт – Анализ данных (рис. 7.2), а также при создании формул становится доступной новая группа функций – статистические.

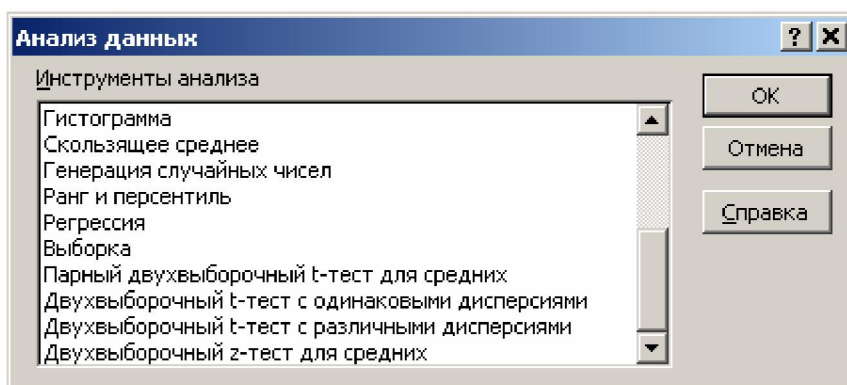


Рис. 7.2. Окно анализа данных, вызываемое из меню «Сервис/Анализ данных...»

В рамках Microsoft Excel с помощью встроенных статистических команд можно провести:

- описательный статистический анализ;
- ранжирование данных;
- графический анализ данных;
- прогнозирование данных;
- регрессионный анализ и др.

В табл. 7.2 приведены в алфавитном порядке некоторые статистические функции, позволяющие пользователю реализовать обработку данных непосредственно на листе электронной таблицы.

Таблица 7.2

Статистические функции пакета Microsoft Excel

Функция	Назначение функции и ее аргументы
1	2
ВЕРОЯТНОСТЬ	Возвращает вероятность того, что значение из интервала находится внутри заданных пределов. Если верхний_предел не задан, то возвращается вероятность того, что значения в аргументе x_интервал равняются значению аргумента нижний_предел. ВЕРОЯТНОСТЬ (x_интервал; интервал_вероятностей; нижний_предел; верхний_предел).
ДИСП	Оценивает дисперсию по выборке ДИСП (число1; число2; ...)
ДИСПР	Вычисляет дисперсию для генеральной совокупности ДИСПР (число1; число2; ...)
ДОВЕРИТ	Возвращает доверительный интервал для среднего генеральной совокупности ДОВЕРИТ (альфа; станд_откл; размер)
КВАДРОТКЛ	Возвращает сумму квадратов отклонений точек данных от их среднего КВАДРОТКЛ (число1; число2;...)
КВПИРСОН	Возвращает квадрат коэффициента корреляции Пирсона для точек данных в аргументах известные_значения_y и известные_значения_x КВПИРСОН (известные_значения_y; известные_значения_x)
КОРРЕЛ	Возвращает коэффициент корреляции между интервалами ячеек массив1 и массив2 КОРРЕЛ (массив1; массив2)

1	2
ЛГРФПРИБЛ	<p>В регрессионном анализе вычисляет экспоненциальную кривую, аппроксимирующую данные, и возвращает массив значений, описывающий эту кривую. Поскольку данная функция возвращает массив значений, она должна вводиться как формула для работы с массивами. Уравнение кривой следующее:</p> $y = b \cdot m^x \text{ или } y = (b \cdot (m_1^{x^1}) \cdot (m_2^{x^2}) \cdot \dots \cdot (m_n^{x^n}))$ <p>(при наличии нескольких значений x), где зависимые значения y являются функцией независимых значений x. Значения m являются основанием для возведения в степень x, а значения b постоянны. Отметим, что y, x и m могут быть векторами. Функция ЛГРФПРИБЛ возвращает массив $\{m_n; m_{n-1}; \dots; m_1; b\}$.</p> <p>ЛГРФПРИБЛ(известные_значения_y; известные_значения_x; конст; статистика)</p>
ЛИНЕЙН	<p>Рассчитывает статистику для ряда с применением метода наименьших квадратов, чтобы вычислить прямую линию, которая наилучшим образом аппроксимирует имеющиеся данные. Функция возвращает массив, который описывает полученную прямую. Поскольку возвращается массив значений, функция должна задаваться в виде формулы массива.</p> <p>ЛИНЕЙН(известные_значения_y; известные_значения_x; конст; статистика)</p>
МАКС	<p>Возвращает наибольшее значение из набора значений</p> <p>МАКС(число1; число2; ...)</p>
МЕДИАНА	<p>Возвращает медиану заданных чисел</p> <p>МЕДИАНА(число1; число2; ...)</p>
МИН	<p>Возвращает наименьшее значение в списке аргументов</p> <p>МИН(число1; число2; ...)</p>
МОДА	<p>Возвращает наиболее часто встречающееся или повторяющееся значение в массиве или интервале данных</p> <p>МОДА(число1; число2; ...)</p>
НАКЛОН	<p>Возвращает наклон линии линейной регрессии для точек данных в аргументах известные_значения_y и известные_значения_x. Наклон определяется как частное от деления расстояния по вертикали на расстояние по горизонтали между двумя любыми точками прямой, то есть наклон – это скорость изменения значений вдоль прямой</p> <p>НАКЛОН(известные_значения_y; известные_значения_x)</p>
НОРМАЛИЗАЦИЯ	<p>Возвращает нормализованное значение для распределения, характеризуемого средним и стандартным отклонением</p> <p>НОРМАЛИЗАЦИЯ(x; среднее; стандартное_откл)</p>

1	2
НОРМОБР	Возвращает обратное нормальное распределение для указанного среднего и стандартного отклонения НОРМОБР (вероятность; среднее; стандартное_откл)
НОРМРАСП	Возвращает значение нормальной функции распределения для указанного среднего и стандартного отклонения НОРМРАСП (x; среднее; стандартное_откл; интегральная)
НОРМСТОБР	Возвращает обратное значение стандартного нормального распределения u НОРМСТОБР (вероятность)
НОРМСТРАСП	Возвращает стандартное нормальное интегральное распределение. Это распределение имеет среднее, равное нулю, и стандартное отклонение, равное единице. Эта функция используется вместо таблицы для стандартной нормальной кривой НОРМСТРАСП (u)
ОТРЕЗОК	Вычисляет точку пересечения линии с осью y , используя известные_значения_ x и известные_значения_ y ОТРЕЗОК (известные_значения_ x ; известные_значения_ y)
ПИРСОН	Возвращает коэффициент корреляции Пирсона r (выборочный коэффициент корреляции), безразмерный индекс в интервале от -1,0 до 1,0 включительно ПИРСОН (массив1; массив2)
СРГЕОМ	Возвращает среднее геометрическое значений массива или интервала положительных чисел СРГЕОМ (число1; число2; ...)
СРЗНАЧ	Возвращает среднее арифметическое своих аргументов СРЗНАЧ (число1; число2; ...)
СРОТКЛ	Среднее абсолютных значений отклонений точек данных от среднего СРОТКЛ (число1; число2; ...)
СТАНДОТКЛОН	Оценивает стандартное отклонение по выборке СТАНДОТКЛОН (число1; число2; ...)
СТАНДОТКЛОНП	Вычисляет стандартное отклонение по генеральной совокупности СТАНДОТКЛОНП (число1; число2; ...)
СТЬЮДРАСП	Возвращает t -распределение Стьюдента СТЬЮДРАСП (x ; степени_свободы; хвосты)
СТЬЮДРАСПОБР	Возвращает обратное распределение Стьюдента для заданного числа степеней свободы СТЬЮДРАСПОБР (вероятность; степени_свободы)

1	2
ТЕНДЕНЦИЯ	Определяет предсказанные значения в соответствии с линейным трендом для заданного массива (методом наименьших квадратов) ТЕНДЕНЦИЯ (известные_значения_y; известные_значения_x; новые_значения_x; конст)
ТТЕСТ	Возвращает вероятность, соответствующую критерию Стьюдента ТТЕСТ (массив1; массив2; хвосты; тип)
ФИШЕР	Возвращает преобразование Фишера для аргумента x ФИШЕР (x)
ФИШЕРОБР	Возвращает обратное преобразование Фишера ФИШЕРОБР (y)
ХИ2ОБР	Возвращает значение обратное к односторонней вероятности распределения χ^2 (хи-квадрат) ХИ2ОБР (вероятность; степени_свободы)
ХИ2РАСП	Возвращает одностороннюю вероятность (P) распределения χ^2 (хи-квадрат, распределения Пирсона) ХИ2РАСП (x; степени_свободы)
ЧАСТОТА	Вычисляет частоту появления значений в интервале значений и возвращает массив цифр ЧАСТОТА (массив_данных; массив_карманов)
ЭКЦЕСС	Возвращает эксцесс множества данных ЭКЦЕСС (число1; число2; ...)
ФРАСП	Возвращает F-распределение вероятности (распределение Фишера) ФРАСП (x; степени_свободы1; степени_свободы2)
ФРАСПОБР	Возвращает обратное значение для F-распределения вероятностей (критерий Фишера) ФРАСПОБР (вероятность; степени_свободы1; степени_свободы2)

7.3. Краткое описание системы STATISTICA

STATISTICA – это универсальная компьютерная интегрированная система, предназначенная для статистического анализа и визуализации данных, а также разработки пользовательских приложений, содержащая широкий набор процедур анализа для применения в научных исследованиях, технике и бизнесе. Отличительной чертой системы STATISTICA является то, что в ней реализован так называемый графически-ориентированный подход к анализу данных. Смысл подхода состоит в том, чтобы получать всестороннее визуальное представление данных на всех эта-

пах статистической обработки и на основе этого представления выбирать следующий шаг анализа. Детальное рассмотрение примеров поможет читателю овладеть основными приемами работы в этой системе и успешно использовать ее в своей инженерной практике. Среди возможностей пакета отметим те задачи, с решением которых приходится сталкиваться в своей практической деятельности инженеру-металлургу: описательные статистики, анализ многомерных таблиц, подгонка распределений, многомерная регрессия, нелинейная регрессия, дисперсионный и ковариационный анализ. Для отображения результатов статистической обработки используются разнообразные графики: гистограммы, диаграммы рассеяния, круговые диаграммы, вероятностные графики, графики поверхностей и многие другие.

Помимо общих статистических и графических средств, в системе имеются специализированные модули, например, для решения инженерно-технических и, в частности, промышленных задач: карты контроля качества, анализ процессов и планирование эксперимента. Работа со всеми модулями происходит в рамках единого программного пакета, для которого можно выбирать один из нескольких предложенных интерфейсов пользователя. С помощью реализованных в системе STATISTICA мощных языков программирования и языка макрокоманд, снабженных специальными средствами поддержки, пользователь может создать законченные статистические модули (процедуры) и встраивать их в различные другие приложения или вычислительные среды.

Программа использует стандартный интерфейс электронных таблиц. Предусмотрены возможность обмена данными со всеми популярными СУБД (в том числе с MS Excel и MS Access), работа с удаленными базами данных, поддержка технологий OLE, DDE и ODBC. Встроенные графический и текстовый редакторы позволяют публиковать полученные материалы на Internet-страницах.

7.3.1. Общая структура системы

Система STATISTICA состоит из следующих основных компонент:

- многофункциональной системы для работы с данными, которая включает в себя электронные таблицы (*Spreadsheet*) для ввода и задания исходных данных, а также специальных таблиц (*Scrollsheet*) для вывода численных результатов анализа. Для сложной (специализированной) обработки данных в STATISTICA имеется модуль *Управление данными*. Для статистической обработки чрезвычайно

больших массивов данных имеется специальный инструмент *Менеджер мега-файлов*, который может быть использован и для предварительной обработки данных перед вводом их непосредственно в электронную таблицу STATISTICA;

- мощной графической системы для визуализации данных и результатов статистического анализа;
- набора статистических модулей, в которых собраны группы логически связанных между собой статистических процедур. В любом конкретном модуле можно выполнить определенный способ статистической обработки, не обращаясь к процедурам из других модулей. Каждый модуль является полноценным WINDOWS приложением. Поэтому пользователь имеет возможность одновременной работы как с одним, так и с несколькими модулями. Переключаться между ними можно как между обычными WINDOWS приложениями, например, при помощи комбинации клавиш [ALT]+[TAB] или выбирая их на панели задач. Все основные операции при работе с данными и графические возможности доступны в любом статистическом модуле и на любом шаге анализа;
- специального инструментария для подготовки отчетов. При помощи текстового редактора, встроенного в систему, можно готовить полноценные отчеты. В STATISTICA также имеется возможность автоматического создания отчетов;
- встроенных языков SCL и STATISTICA BASIC, которые позволяют автоматизировать рутинные процессы обработки данных в системе.

Все структурные компоненты STATISTICA настолько тесно интегрированы между собой, что разделение на различные компоненты во многом условно и является полезным лишь для изучения системы с методической точки зрения.

Система STATISTICA работает с четырьмя различными типами документов, которые соответствуют основным структурным компонентам системы:

- электронная таблица *Spreadsheet*, которая предназначена для ввода исходных данных и их преобразования;
- электронная таблица *Scrollsheet* для вывода численных и текстовых результатов анализа;
- график – документ в специальном графическом формате для визуализации графического представления численной информации;
- отчет – документ в формате *RTF (Расширенный текстовый формат)* для вывода текстовой и графической информации.

В соответствии со стандартами среды WINDOWS каждый тип документа выводится в своем собственном окне в рабочей области системы STATISTICA. Как только это окно становится активным, изменяется панель инструментов и меню. В них появляются команды и кнопки, доступные для активного документа. Имеется несколько различных способов работы с системой STATISTICA.

7.3.2. Возможные способы взаимодействия с системой

Статистический анализ данных может быть проведен пользователем в одном из следующих режимов.

Интерактивный режим работы. В этом случае взаимодействие с системой осуществляется при помощи последовательного выбора различных команд из меню. Этот способ работы применяется обычно на этапе предварительного анализа данных. Интерактивный режим работы с системой удобен на этапе выбора математической модели явления и метода статистического анализа. После того как выбор сделан, рекомендуется использовать для автоматизации выполнения рутинных задач обработки специальные макрокоманды различных типов и встроенные в систему языки (SCL и STATISTICA BASIC).

Использование макрокоманд. В системе STATISTICA имеется возможность записи последовательности команд в одну макрокоманду. При этом можно записывать как последовательности нажатий клавиш на клавиатуре, так и движения мыши. Это удобное средство позволяет автоматизировать выполнение часто повторяющихся шагов статистического анализа.

При помощи встроенного командного языка системы STATISTICA (язык SCL – STATISTICA Command Language) пользователь имеет возможность выполнять статистическую обработку данных в пакетном режиме.

При помощи встроенного процедурного языка STATISTICA BASIC пользователь может написать свои собственные процедуры обработки данных. Это мощный язык, ориентированный на структуру данных системы STATISTICA, содержит большое количество специальных математических и статистических функций (например, вычисление всевозможных статистических распределений и т.д.).

7.3.3. Ввод данных

Данные в STATISTICA организованы в виде электронной таблицы – *Spreadsheet*. Они могут содержать как численную, так и текстовую информацию. Данные в электронной таблице могут иметь различные форматы (например, даты, времени и др.). Электронные таблицы в STATISTICA поддерживают различные типы операций с данными, такие как: операции с использованием буфера обмена WINDOWS, операции с выделенными блоками значений (аналогично MS Excel).

Ввести данные в электронную таблицу можно одним из следующих способов.

Непосредственно ввести их в электронную таблицу с клавиатуры. Для автоматизации ручного ввода данных в STATISTICA имеются развитые инструментальные средства.

Вычислить новые данные на основе уже введенных данных при помощи формул, которые можно задать в электронной таблице. При этом имеется возможность быстрого доступа к большому количеству специализированных статистических функций, допускается использование логических операторов.

Воспользоваться данными, подготовленными в другом приложении. При этом доступны следующие способы ввода данных из других приложений STATISTICA:

- операции копирования данных через буфер обмена WINDOWS;
- импорт данных из наиболее популярных приложений, например электронных таблиц, систем управления базами данных и др.;
- использование механизма динамической связи DDE между данными в STATISTICA и другим WINDOWS приложением. В этом случае все изменения данных, внесенные позднее в источник (WINDOWS приложение), будут автоматически (динамически) отражены в файле исходных данных для системы STATISTICA.

На любом этапе ввода данных система STATISTICA позволяет быстро вычислить основные статистические характеристики данных, отобразить их графически и перейти к статистическому анализу.

7.3.4. Вывод численных и текстовых результатов анализа

Численные результаты статистического анализа в системе STATISTICA выводятся в виде специальных электронных таблиц, которые называются таблицами вывода результатов – *Scrollsheet*. Таблицы *Scrollsheet* могут содержать любую информацию (как численную, так и текстовую) размером от короткой строчки до нескольких мегабайтов.

STATISTICA содержит большое количество инструментов для просмотра результатов статистического анализа и их визуализации. Они включают в себя стандартные операции по редактированию таблицы (вставка, удаление, операции над блоками, автозаполнение блоков и др.), операции просмотра (подвижные границы столбцов, разделение прокрутки в таблице и др.), доступ к основным статистикам и графическим возможностям системы STATISTICA. При выводе целого ряда результатов (например, корреляционной матрицы) STATISTICA отмечает значимые коэффициенты корреляции цветом. Пользователь имеет возможность выделить при помощи цвета необходимые значения в таблице *Scrollsheet*.

Если пользователю необходимо провести детальный статистический анализ промежуточных результатов, то можно сохранить таблицу *Scrollsheet* в формате файла STATISTICA и далее работать с ним, как с обычными данными.

Кроме вывода результатов анализа в виде отдельных окон с графиками и таблицами *Scrollsheet*, в системе имеется возможность создания отчета, в окно которого может быть выведена вся эта информация – это документ (в формате RTF), который может содержать любую текстовую или графическую информацию. В STATISTICA имеется возможность автоматического создания отчета, так называемого автоотчета. При этом любая таблица *Scrollsheet* или график могут автоматически быть направлены в отчет.

7.3.5. Статистические процедуры системы STATISTICA

Статистические процедуры системы STATISTICA, как уже было отмечено, сгруппированы в нескольких специализированных статистических модулях (рис. 7.3). В каждом модуле можно выполнить определенный способ обработки, не обращаясь к процедурам из других модулей. Ниже приводится краткое описание отдельных статистических модулей, знакомство с которыми поможет инженеру-исследователю в

его практике для профессиональной обработки, анализа и представления результатов эксперимента (промышленного, лабораторного, вычислительного и др.).

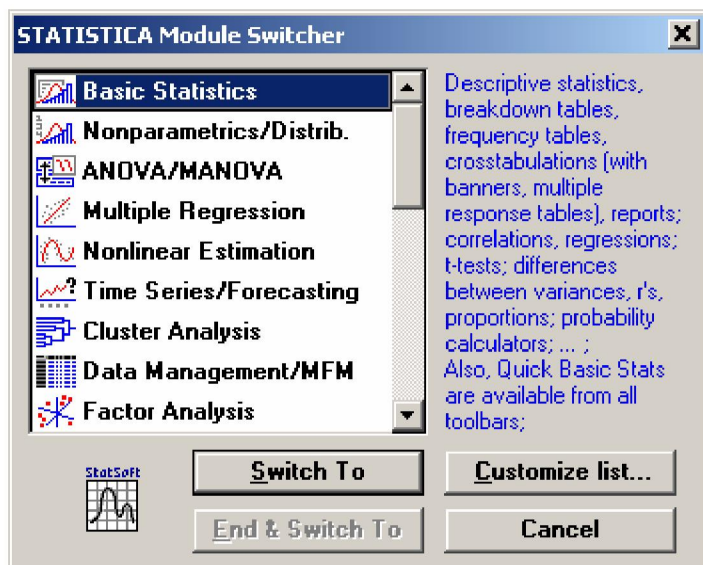


Рис. 7.3. Основные статистические модули системы STATISTICA

Модуль Основные статистики и таблицы (Basic Statistics/Tables) включает в себя следующие группы статистических процедур:

Описательные статистики (Descriptive statistics). Группа производит вычисление практически всех описательных статистик, среднее арифметическое, выборочную дисперсию, стандартное отклонение, медиану, моду, максимальное и минимальное значения, размах, доверительные интервалы для среднего и многие другие описательные статистики. Здесь же предусмотрен широкий выбор критериев для тестирования нормальности распределения. Практически все описательные статистики могут быть вычислены для данных, разделенных на группы с помощью одной или нескольких группирующих переменных. Имеется возможность интерактивного удаления выбросов на графике, выделение и маркировка необходимых подмножеств на графике, сглаживание данных и другие возможности.

Корреляционные матрицы (Correlation matrices). Данная группа включает большое количество средств, позволяющих исследовать зависимости между переменными путем вычисления практически всех общих мер зависимости (коэффициентов корреляции).

t-критерии для зависимых и независимых выборок (t-test for independent and dependent samples). Эта группа процедур позволяет осуществить проверку истинности статистических гипотез относительно наблюдаемых случайных величин путем выполнения специальных тестов.

Таблицы частот (Frequency tables). Группа позволяют строить таблицы частот и гистограммы выбранных переменных. При этом значения переменных можно разбивать на классы и группировать произвольным образом.

Калькулятор вероятностных распределений (Probability calculator). Данная группа позволяет вычислить характеристики многих стандартных вероятностных распределений: нормального, хи-квадрат, Стьюдента, F-распределения и др.

Модуль Множественная регрессия (Multiple regression) включает в себя исчерпывающий набор средств множественной линейной и фиксированной нелинейной (в частности, полиномиальной, экспоненциальной, логарифмической и др.) регрессий, позволяющих вычислять неизвестные коэффициенты в заранее заданных пользователем регрессионных моделях.

Модуль Нелинейное оценивание (Nonlinear estimation) дает возможность оценить практически любые определенные пользователем нелинейные модели, осуществить подгонку к наблюдаемым данным кривой, по существу, любого типа. Важным преимуществом данного модуля, в отличие от других программ нелинейного оценивания, является то, что в нем не накладывается ограничения на размер обрабатываемого файла данных. Оценки коэффициентов нелинейной модели могут быть построены с помощью оценок метода наименьших квадратов, метода максимального правдоподобия или заданной пользователем функции потерь. Пользователь может выбрать одну из нескольких вычислительных процедур: квазиньютоновский метод, симплекс-метод и др. Кроме того, пользователь может сам определить любой тип нелинейной модели, набрав соответствующее уравнение в редакторе системы.

Модуль Дисперсионный анализ (ANOVA/MANOVA) дает возможность оценить степень воздействия различных факторов на измеряемые данные и выделить среди них наиболее значимые (существенные). Для проверки основных предположений дисперсионного анализа имеется широкий выбор статистических процедур, в частности критерии Фишера, Кохрена, Бартлета и др.

Модуль Факторный анализ (Factor analysis) позволяет проводить факторный анализ, основная цель которого заключается в том, чтобы выделить скрытые общие факторы, т.е. воздействующие на все параметры объекта, а не на какой-то один параметр или группу. Выделяемые общие факторы определяют связи между наблюдаемыми параметрами объекта.

Модуль Непараметрическая статистика и подгонка распределения (Nonparametrics/Distribution) дает возможность сравнить распределение наблюдаемых величин с большим количеством различных теоретических распределений. Имеется возможность подогнать к данным нормальное, логнормальное, экспоненци-

альное, хи-квадрат, пуассоновское и др. распределения. Точность подгонки оценивается с помощью различных критериев (хи-квадрат, Колмогорова–Смирнова и др.).

Модуль Анализ временных рядов и прогнозирование (Time Series/Forecasting) позволяет строить модель, описывающую ряд данных, сгладить его, спрогнозировать будущие значения временного ряда на основе наблюдаемых до данного момента, построить регрессионные зависимости одного ряда от другого, провести спектральный или Фурье-анализ ряда и т.д. Модуль также включает процедуры автокорреляционного анализа.

Возможности системы STATISTICA далеко не исчерпываются перечисленными выше модулями. Кроме них, система содержит ряд модулей, предназначенных для более детального статистического анализа данных и необходимость в которых на практике возникает крайне редко.

7.3.6. Структура диалога пользователя в системе STATISTICA

Структура диалога пользователя в каждом статистическом модуле имеет общие черты:

- 1) после выбора из переключателя модулей (рис. 7.3) открывается стартовая панель выбранного статистического модуля;
- 2) далее необходимо открыть файл данных и выбрать переменные для анализа из открытого файла;
- 3) затем выбираются метод анализа и конкретная вычислительная процедура с соответствующими параметрами расчета из меню в стартовой панели модуля;
- 4) далее запускается вычислительная процедура. Если процедура итерационная, то система дает возможность на каждом шаге просмотреть результаты в появившемся на экране окне и при необходимости добавить число итераций для увеличения точности оценок;
- 5) используя графические возможности и специальные таблицы вывода с вычисленными разнообразными статистиками, можно осуществить всесторонний просмотр и анализ результатов;
- 6) выбирается следующий шаг анализа.

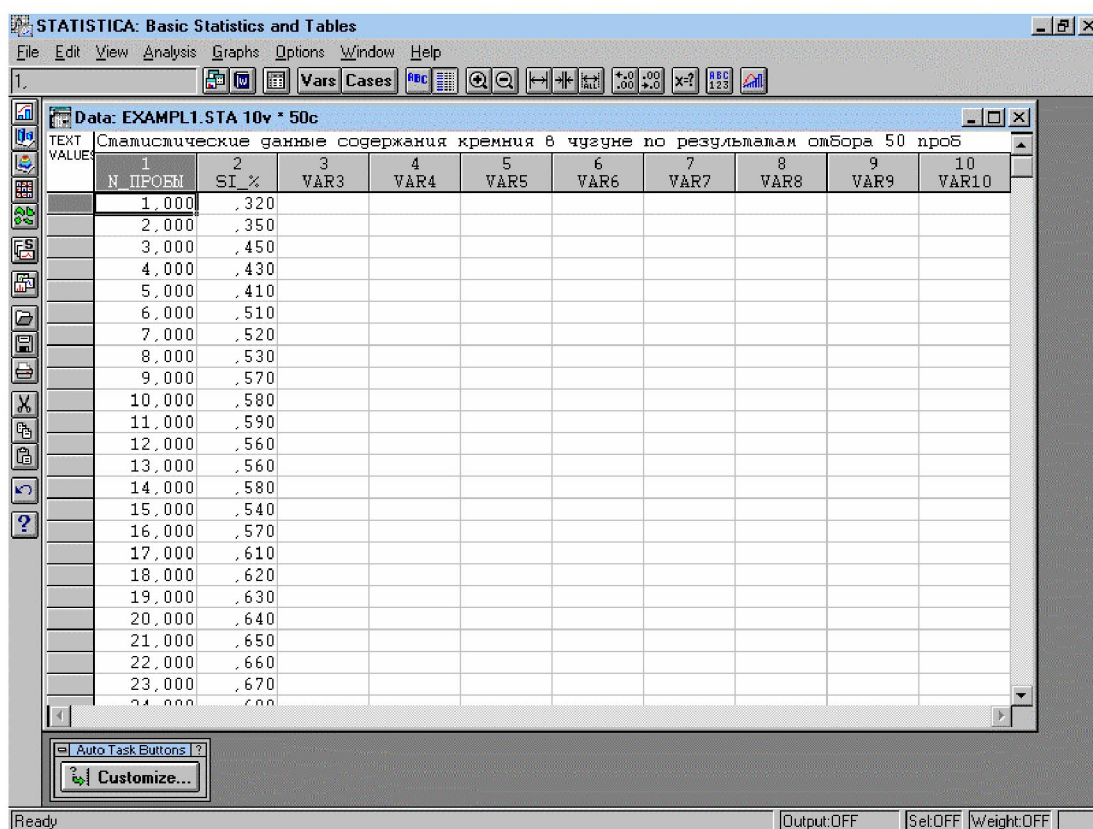
В сложном проекте следует работать с различными модулями, последовательно переключаясь между ними.

7.3.7. Примеры использования системы STATISTICA

Расчет основных характеристик случайных величин

Запускаем систему STATISTICA и выбираем статистический модуль Basic Statistics/Tables (*Основные статистики и таблицы*). Создаем новый файл исходных данных, выбирая из меню пункт File/New data, и присваиваем ему произвольное имя, например exampl1.sta.

Исходные данные для анализа возьмем из примера 3.8, в котором приведены результаты обработки 50 проб передельного чугуна на предмет содержания в них кремния (табл. 3.7). Заполним таблицу в системе STATISTICA с исходными данными, как показано на рис. 7.4. Для этого создаем две переменные: первая содержит номер пробы чугуна (N_ПРОБЫ), вторая – процентное содержание кремния в чугуне для соответствующей пробы (SI_%). Размер таблицы в системе по умолчанию принят 10 на 10 (10 переменных с именами VAR1, VAR2, ..., VAR10 и 10 случаев). Чтобы изменить имя переменной, необходимо выбрать пункт меню Edit/Variables/Current Specs или нажать комбинацию клавиш [Ctrl]+[F2], а затем в диалоге указать нужное имя.



1	2	3	4	5	6	7	8	9	10
N_ПРОБЫ	SI_%	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10
1,000	,320								
2,000	,350								
3,000	,450								
4,000	,430								
5,000	,410								
6,000	,510								
7,000	,520								
8,000	,530								
9,000	,570								
10,000	,580								
11,000	,590								
12,000	,560								
13,000	,560								
14,000	,580								
15,000	,540								
16,000	,570								
17,000	,610								
18,000	,620								
19,000	,630								
20,000	,640								
21,000	,650								
22,000	,660								
23,000	,670								
24,000	,680								

Рис.7.4. В рабочем окне системы STATISTICA расположен файл с исходными данными содержания кремния в чугуне

После того как подготовлен файл исходных данных, выбираем пункт меню Analysis/Descriptive Statistics (*Описательная статистика*). В появившемся диалоговом окне, вид которого показан на рис. 7.5, следует выбрать переменную для анализа нажатием кнопки Variables. Мы выбрали переменную, содержащую данные о процентном содержании кремния в чугуна, имя выбранной переменной отражается рядом с кнопкой Variables. Диалог *Описательная статистика* позволяет:

- вычислить разнообразные выборочные характеристики (Statistics): Median&quartiles – медиана и квартили, Conf. Limits for means – доверительные границы для среднего. В строке Alpha error можно задать требуемый уровень значимости. Напомним, что уровень значимости – это вероятность неправильного отвержения гипотезы, когда она верна (подробнее см. п. 3.3). Более расширенный набор статистик для расчета можно выбрать нажатием кнопки More statistics (рис. 7.6): Mean – среднее арифметическое, Sum – сумма, Median – медиана, Standard Deviation – стандартное отклонение, Variance – дисперсия, Minimum&maximum – минимум и максимум, Range – размах, т.е. разность между максимумом и минимумом, Skewness – коэффициент асимметрии, Kurtosis – коэффициент эксцесса:

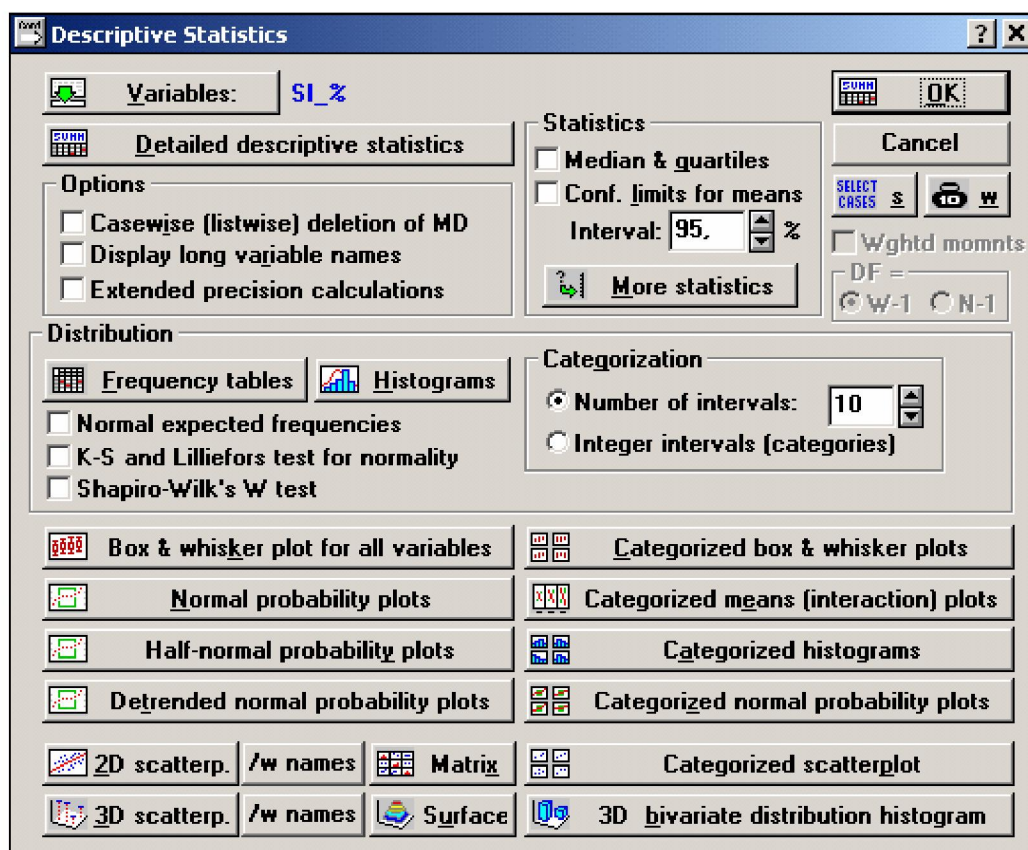


Рис.7.5. Вид экрана меню Descriptive Statistics (*Описательные статистика*)

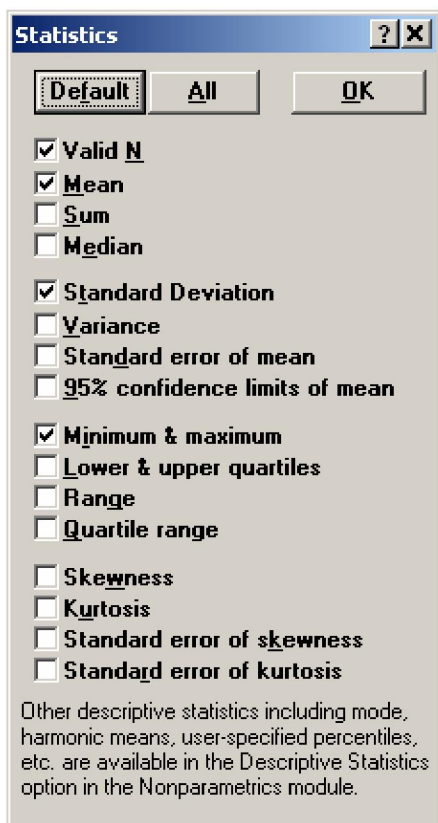


Рис.7.6. Меню Статистики

- построить для выборки таблицу частот (Frequency Tables) и гистограмму частот (Histograms); отметив пункт Normal expected frequencies, можно нанести на гистограмму кривую нормального распределения и визуально оценить соответствие исходных данных нормальному закону распределения;
- проверить гипотезу о нормальности распределения наблюдаемых случайных величин с использованием критерия Колмогорова–Смирнова, выбирая его в разделе Distribution.

Для визуализации результатов имеется возможность построения разнообразных графиков, вызываемых нажатием соответствующей кнопки в нижней части экрана.

Результаты статистического анализа выводятся в специальное окно. Для данных из примера вид окна с результатами показан на рис. 7.7, из которых следует, что среднее арифметическое (математическое ожидание) содержания кремния в чугуне составляет 0,6504%; выборочная дисперсия 0,0185%; максимальное и минимальное значения равны соответственно 0,32 и 0,95%; действительное содержание кремния в чугуне с вероятностью 95% лежит в интервале от 0,6117 до 0,6891%. Заметим, что эти данные близки к результатам, полученным ранее в примере 3.8 с помощью пакета Microsoft Excel, небольшие расхождения объясняются точностью представления результатов.

Continue...	Valid N	Mean	Confid. -95,000%	Confid. +95,000%	Median	Minimum	Maximum	Variance	Std.Dev.
SI %	50	,6504	,6117	,6891	,6500	,3200	,9500	,0185	,1361

Рис.7.7. Вид окна с результатами расчета статистик из примера 3.8

Проверка нормальности распределения

Проверим гипотезу о нормальности распределения данных из примера. Для этого в окне диалога *Описательные статистики*, показанного на рис. 7.5, отметим мышью пункт *K-S test and Lilliefors test normality (Тест Колмогорова–Смирнова для проверки нормальности)* и нажмем кнопку *Histograms*. В результате на экране появится окно (рис. 7.8), на котором изображена гистограмма переменной *SI_%* и дополнительно нанесена линия нормального распределения. В верхней части гистограммы показан рассчитанный параметр теста Колмогорова – Смирнова – критерий *D* (см. п. 3.6).

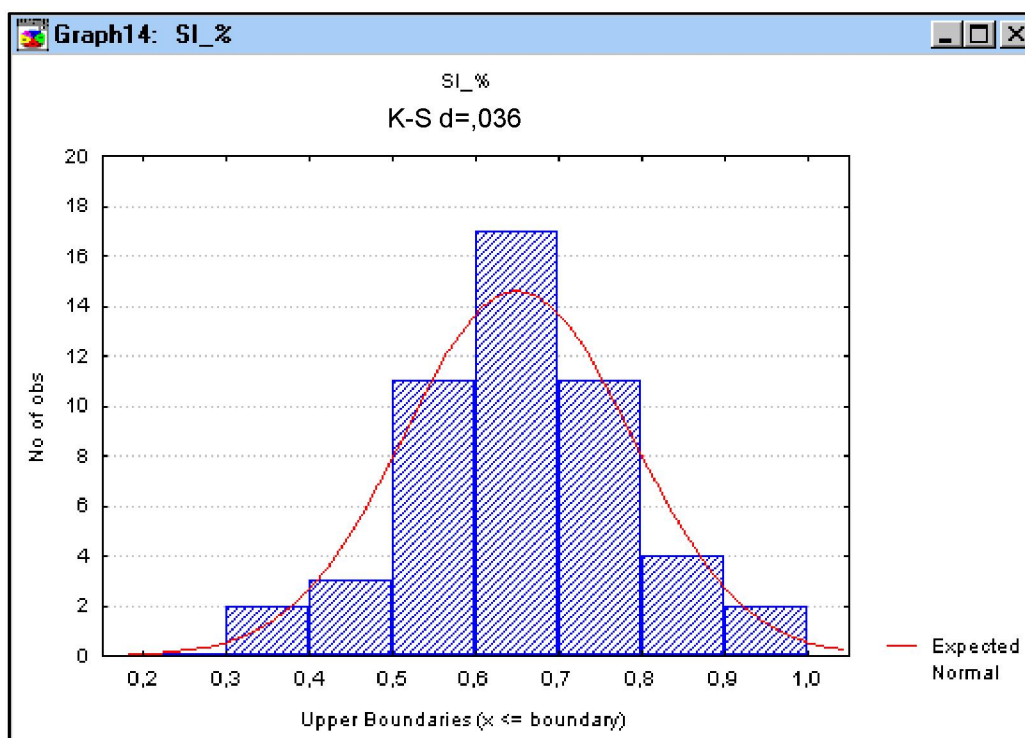


Рис.7.8. Гистограмма распределения содержания кремния в чугуна с результатами проверки гипотезы о нормальности распределения

Напомним, что критерий *D* называется критерием согласия, поскольку он проверяет, в какой степени наблюдаемые значения случайной величины согласуются с функцией нормального распределения: чем меньше величина *D*, тем в меньшей степени эмпирическая функция распределения случайной величины отличается от нормальной функции распределения. В нашем случае критерий $D=0,07934$, следовательно, гипотеза о нормальности распределения данных о содержании кремния в

чугуне подтверждается. В противном случае система выдала бы соответствующее сообщение и выделила бы данный критерий отличительным цветом.

Регрессионный анализ

Применение системы STATISTICA для регрессионного анализа рассмотрим на примере исследования взаимосвязи среднемесячного удельного расхода кокса и соответствующей величины удельного выхода шлака по данным работы одной из доменных печей ОАО "Магнитогорский металлургический комбинат" (ОАО ММК), которые отражены в табл. 7.3. Напомним, что задача регрессионного анализа состоит в том, чтобы по наблюдениям входных (X) и выходных (Y) параметров:

- построить регрессионную модель (уравнение регрессии), т.е. оценить параметры модели (коэффициенты уравнения) наилучшим образом;
- построить доверительные интервалы для коэффициентов модели;
- проверить гипотезу о значимости регрессии;
- оценить степень адекватности модели и т.д.

Из теории доменного процесса известно, что величина расхода кокса зависит от выхода шлака, а не наоборот. Поэтому зависимой переменной Y будет являться величина удельного расхода кокса, а независимой переменной X – величина удельного выхода шлака. Регрессионный анализ будем проводить в несколько этапов.

Таблица 7.3

Фактические данные о расходе кокса (K) и выходе шлака (Ш) на одной из доменных печей ОАО ММК

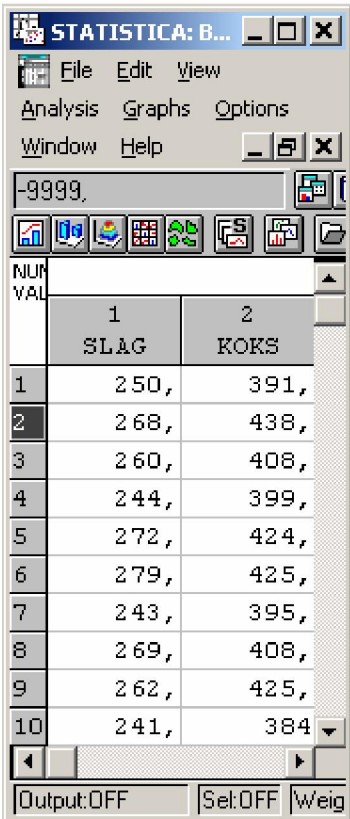
№ п/п	K, кг/т чугуна	Ш, кг/т чугуна	№ п/п	K, кг/т чугуна	Ш, кг/т чугуна	№ п/п	K, кг/т чугуна	Ш, кг/т чугуна
1	2	3	1	2	3	1	2	3
1	391	250	31	403	241	61	393	251
2	438	268	32	440	277	62	418	252
3	408	260	33	416	253	63	426	270
4	399	244	34	445	294	64	412	260
5	424	272	35	384	249	65	419	277
6	425	279	36	435	263	66	433	274

Окончание табл. 7.3

1	2	3	1	2	3	1	2	3
7	395	243	37	389	254	67	387	245
8	408	269	38	406	252	68	451	291
9	425	262	39	384	238	69	398	253
10	384	241	40	399	240	70	403	244
11	424	267	41	419	256	71	425	259
12	403	239	42	418	261	72	382	225
13	409	242	43	398	262	73	397	244
14	416	270	44	421	270	74	449	280
15	410	254	45	398	257	75	385	256
16	392	256	46	398	250	76	385	228
17	403	244	47	438	271	77	386	218
18	377	239	48	437	280	78	418	253
19	419	262	49	411	264	79	408	256
20	382	251	50	406	250	80	442	270
21	430	267	51	409	269	81	406	243
22	406	258	52	391	242	82	413	259
23	379	240	53	399	252	83	436	266
24	420	251	54	435	273	84	384	242
25	389	256	55	420	260	85	429	275
26	411	253	56	416	261	86	392	237
27	419	275	57	420	276	87	390	259
28	413	272	58	386	242	88	427	262
29	392	235	59	390	220	89	398	242
30	418	257	60	427	258	90	445	280

1. Воспользуемся статистическим модулем Nonlinear estimation (*Нелинейное оценивание*), в котором создадим новый файл exampl2.sta и занесем в него данные из табл. 7.3. На рис. 7.9 показан файл с исходными данными. Переменные, содержащие данные об удельных расходе кокса и выходе шлака, обозначены соответственно KOKS и SLAG.

2. Последовательность регрессионного анализа в системе STATISTICA лучше



The screenshot shows the STATISTICA software window with a data table. The table has two columns: 'SLAG' and 'KOKS'. The rows are numbered 1 to 10. The data values are as follows:

NUM	1	2
VAL	SLAG	KOKS
1	250,	391,
2	268,	438,
3	260,	408,
4	244,	399,
5	272,	424,
6	279,	425,
7	243,	395,
8	269,	408,
9	262,	425,
10	241,	384

Рис. 7.9. Файл исходных данных для регрессионного анализа

всего начинать с визуальной оценки положения данных при помощи различных графических средств. Поскольку в нашем случае изучается зависимость типа $Y=f(X)$, то для этой цели подойдет двумерный график Scatterplots из меню Graphs/Stats 2D Graphs. В появившемся диалоговом окне (рис. 7.10) выбираем переменные X и Y нажатием кнопки Variables. Имена выбранных переменных Var X (SLAG) и Var Y (KOKS) отображаются в окне диалога. В этом окне дополнительно можно отметить тип графика (Graph Type) – Regular, модель оценивания (FIT) – первоначально выберем Off, стиль графика (Style) – Normal, величину доверительного интервала и т.д. После нажатия кнопки ОК на экране в отдельном окне появляется построенный график.

3. Из анализа наблюдений положения данных на графике делаем вывод о пригодности для оценивания полиномиальной регрессионной модели. Отметим в диалоговом окне (рис. 7.10) модель оценивания Polynomial и нажмем кнопку ОК. В результате появится отдельное окно с графиком, в котором на точечные данные нанесена кривая, подобранная по методу наименьших квадратов и описываемая многочленом 5-го порядка (рис. 7.11). Уравнение многочлена представлено в заголовке графика и имеет следующий вид:

$$y = -73616,8 + 1508,463 \cdot x - 12,187 \cdot x^2 + 0,049 \cdot x^3 - 9,678 \cdot 10^{-5} \cdot x^4 + 7,619 \cdot 10^{-8} \cdot x^5.$$

Абсолютная величина каждого коэффициента в уравнении регрессии характеризует вклад соответствующей степенной составляющей на параметр отклика y .

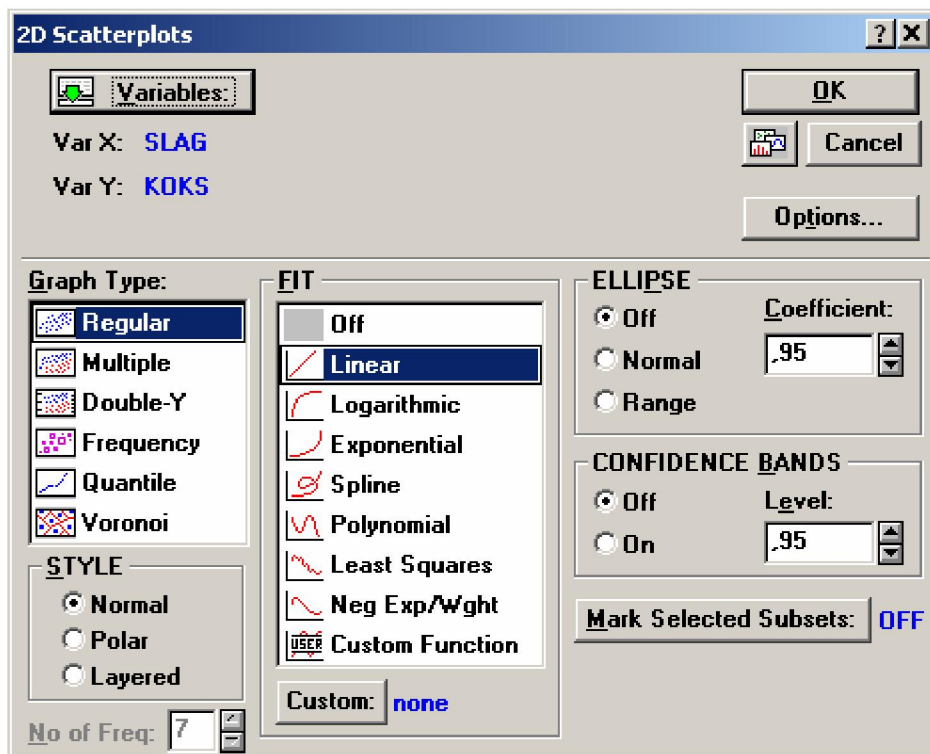


Рис.7.10. Окно построения графиков для подгонки модели оценивания

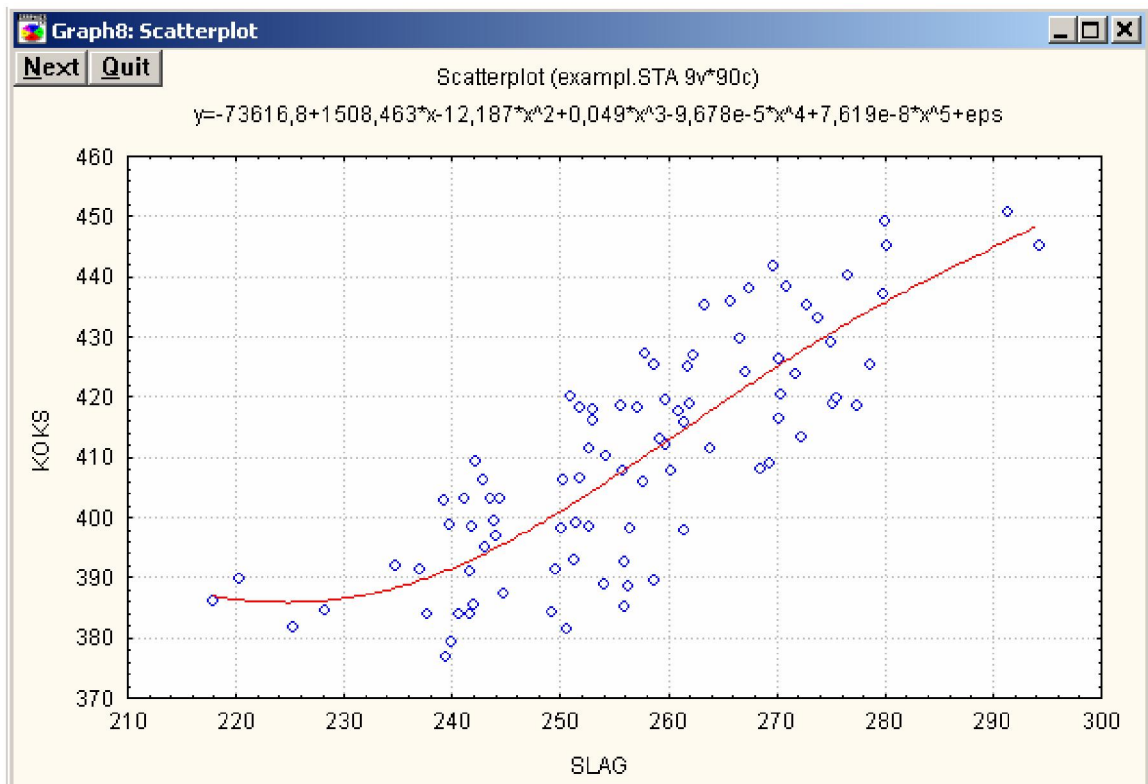


Рис.7.11. Полиномиальная кривая 5-го порядка, рассчитанная по методу наименьших квадратов

4. Для того чтобы проанализировать регрессионную модель, из главного меню системы Analysis (Анализ) выбираем пункт User-specified regression (Определяемая пользователем регрессия). На экране появляется начальное диалоговое окно (рис. 7.12), в котором нажатием кнопки function to be estimated & loss function (функция оценивания и функция потерь) можно с помощью формул задать функцию, которую необходимо оценить, а также определить функцию потерь.

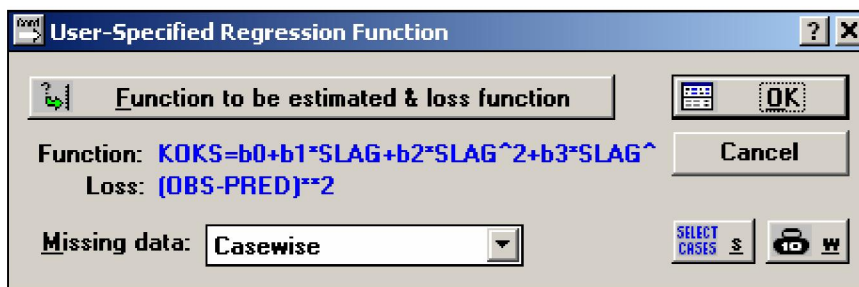


Рис. 7.12. Начальное окно диалога по оценке параметров определяемой пользователем регрессионной кривой

Рассмотрим в качестве примера в дальнейшем зависимость между удельным расходом кокса (KOKS) и удельным выходом шлака (SLAG) в виде полинома 3-й степени:

$$KOKS=b_0+b_1*SLAG+b_2*SLAG^2+ b_3*SLAG^3.$$

Функция потерь по умолчанию задается в виде квадрата отклонения наблюдаемых от предсказанных с помощью регрессионной модели значений $(OBS-PRED)^2$.

5. Нажатие клавиши ОК приводит к появлению окна Model Estimation (Оценивание модели) для выбора метода и начальных установок для пользовательской регрессии (рис. 7.13). В качестве метода оценивания выберем квазиньютоновский. В методах нелинейного оценивания важно правильно подобрать начальные приближения. Неизвестными параметрами модели являются коэффициенты b_0 , b_1 , b_2 и b_3 . Нажав кнопку Start values (Начальные значения), в появившемся диалоговом окне введем начальные значения, предсказанные на основе предварительного графического анализа данных: $b_0=-73616$, $b_1=1508$, $b_2=-12$ и $b_3=0,049$. Нажатие клавиши ОК приводит к появлению окна оценок параметров модели на каждом шаге итерации. После того как оценивание завершится, внизу окна появится сообщение Parameter estimation process converged (Процесс оценивания параметров сошелся).

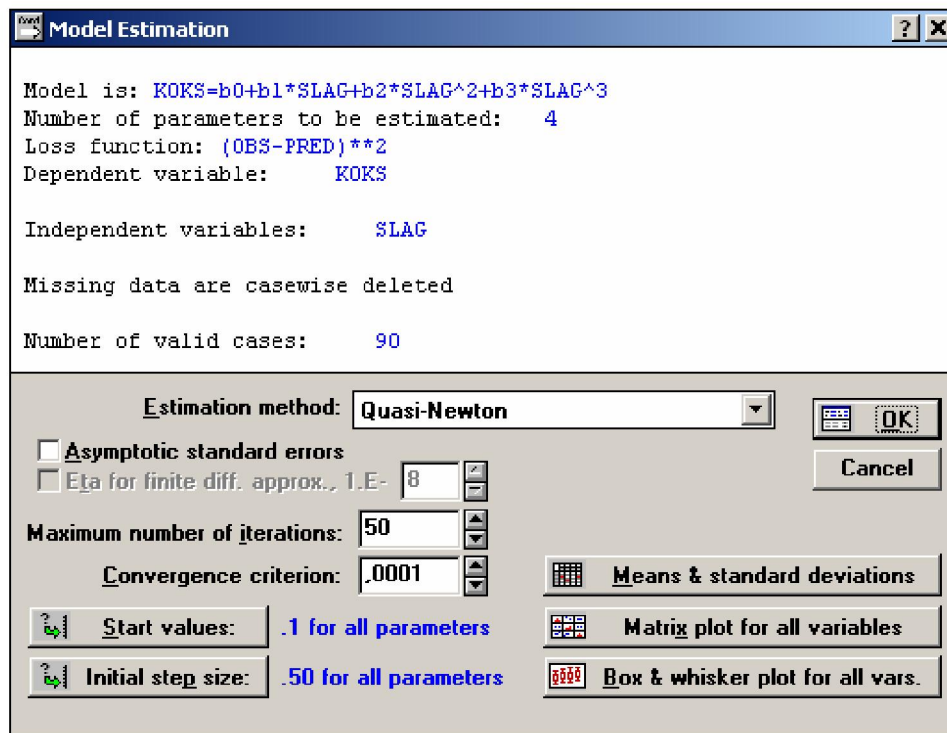


Рис. 7.13. Окно выбора метода и начальных установок для пользовательской регрессии

6. Далее нажимаем кнопку ОК, после чего открывается окно Results (Результаты), показанное на рис. 7.14. Окно результатов имеет следующую структуру: верхняя часть окна – информационная, нижняя содержит функциональные кнопки, позволяющие всесторонне просмотреть результаты анализа.

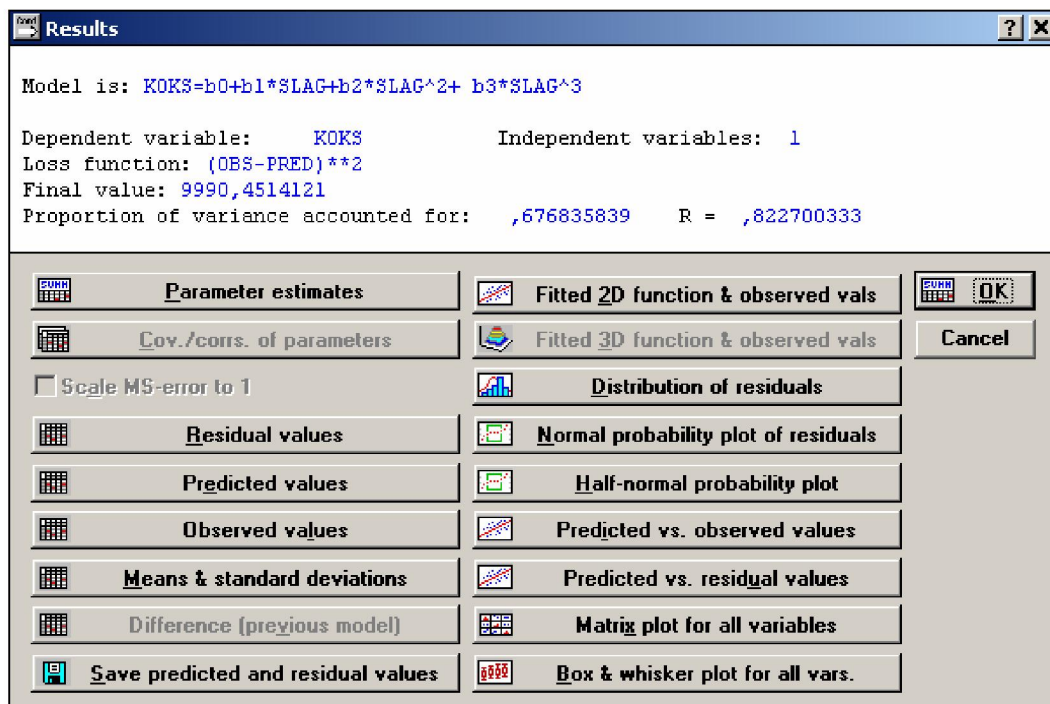


Рис. 7.14. Окно результатов оценивания параметров пользовательской регрессии

Информационная часть содержит краткую информацию о проведенном анализе, а именно:

- Model is – вид модели оценивания. В нашем случае $KOKS=b_0+b_1*SLAG+b_2*SLAG^2+ b_3*SLAG^3$;
- Dependent variable – зависимая переменная. В нашем примере это удельный расход кокса (KOKS);
- Independent variables – количество независимых переменных. В примере независимая переменная одна – удельный выход шлака;
- Loss function – вид функции потерь;
- Final value – последнее значение параметра, по которому система проводила подгонку модели.

Функциональные кнопки позволяют представить результаты в виде таблиц и графиков. Выберем кнопку Parameter estimates (*Параметры оценивания*), и на экране появится окно, в котором отражены численные значения коэффициентов модели (рис. 7.15). Нажатие кнопки Fitted 2D function & observed vals (*Подогнанная функция и наблюдаемые значения*) выводит на экран график результирующей кривой, наложенной на наблюдаемые значения исходных данных (рис. 7.16). Таким образом, уравнение регрессионной модели окончательно примет вид

$$y = 4984,803 - 54,333 \cdot x + 0,21035 \cdot x^2 - 0,000265 \cdot x^3.$$

7. Далее следует оценить поведение остатков (residuals) модели, т.е. разностей между исходными (наблюдаемыми) значениями зависимой переменной и предсказанными с помощью модели. Исследуя остатки модели, можно оценить степень ее адекватности. С помощью функциональных кнопок в данном окне (см. рис. 7.14) можно проанализировать остатки как в графическом виде, так и в электронных таблицах.

Model: KOKS=b0+b1*SLAG+b2*SLAG^2+ b3*SLAG^3 (exempl.sta)				
Continue...	Dep. var: KOKS Loss: (OBS-PRED)**2 Final loss: 9990,4514121 R=,82270 Variance explained: 67,684%			
N=90	B0	B1	B2	B3
Estimate	4984,803	-54,3330	,210350	-,000265

Рис.7.15. Результаты расчета коэффициентов регрессионной модели

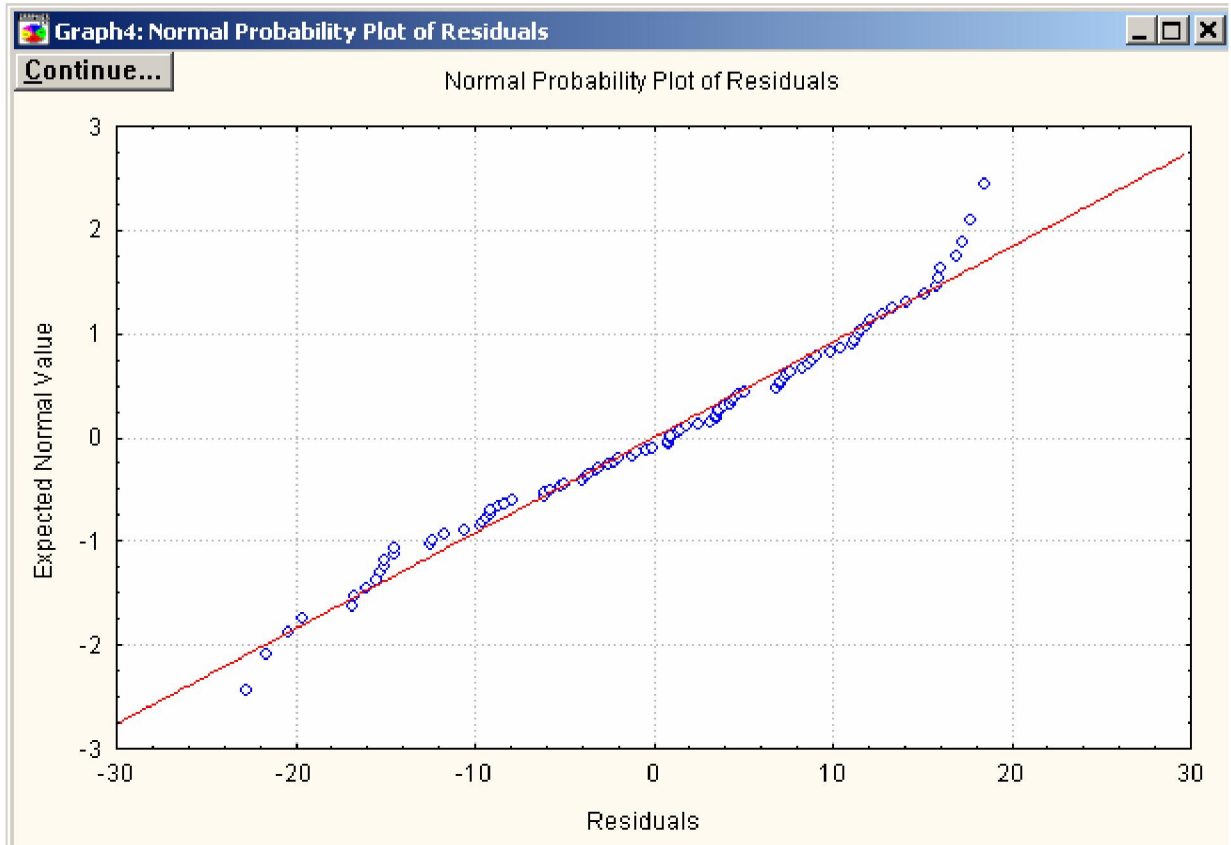


Рис.7.17. График остатков на нормальной вероятностной бумаге

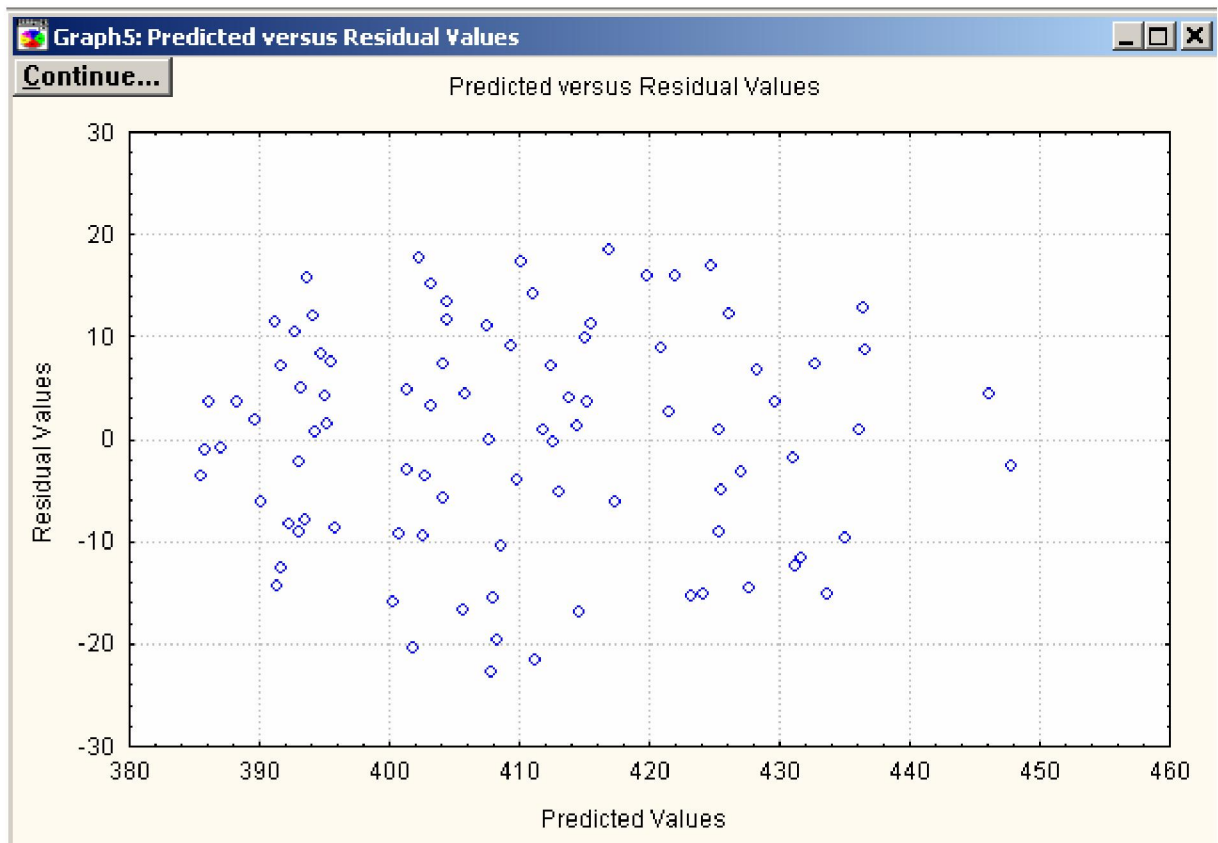


Рис.7.18. Распределение остатков на плоскости