**10. ОБРАБОТКА И АНАЛИЗ РЕЗУЛЬТАТОВ МОДЕЛИРОВАНИЯ**

Решения, принимаемые исследователем по результатам имитационного моделировании, могут быть конструктивными только при выполнении двух основных условий:

* полученные результаты обладают требуемой точностью и достоверностью;
* исследователь способен правильно интерпретировать полученные результаты.

Возможность выполнения первого условия закладывается, в основном, еще на этапе разработки модели и частично – на этапе планирования эксперимента. Достоверность результатов моделирования предполагает, что модель, с помощью которой они получены, не только является «правильной», но и отвечает и некоторым дополнительным требованиям, предъявляемым к имитационным моделям. Эти требования и методы оценки соответствия их созданной модели рассматриваются ниже.

Способность исследователя правильно интерпретировать полученные результаты и принимать на их основе важные решения существенно зависит от степени соответствия формы представления результатов целям моделирования.

Если разработчик модели уверен, что полученные результаты будут использоваться в соответствии с одной, четко определенной целью, форма их представления может быть определена заранее. В этом случае преобразование экспериментальных данных к требуемому виду может производиться либо в ходе эксперимента, либо сразу после его завершения.

Если же заранее конкретизировать цель моделирования сложно или целей несколько, данные должны накапливаться в базе данных и затем уже выдаваться в требуемой форме по запросу пользователя. Как правило, по такому принципу строятся системы автоматизации моделирования.

При правильной организации обработки экспериментальных данных могут быть получены дополнительные сведения о моделируемой системе.

**10.1 Оценка качества имитационной модели**

Оценка качества модели является завершающим этапом ее разработки и преследует две цели:

* проверить соответствие модели ее предназначению (целям исследования);
* оценить достоверность и статистические характеристики результатов, получаемых при проведении модельных экспериментов.

При аналитическом моделировании достоверность результатов определяется двумя основными факторами:

* корректным выбором математического аппарата, используемого для описания исследуемой системы;
* методической ошибкой, присущей данному математическому методу.

При имитационном моделировании на достоверность результатов влияет целый ряд дополнительных факторов, основными из которых являются:

* моделирование случайных факторов, основанное на использовании датчиков СЧ, которые могут вносить «искажения» в поведение модели;
* наличие нестационарного режима работы модели;
* использование нескольких разнотипных математических методов в рамках одной модели;
* зависимость результатов моделирования от плана экспериментов;
* необходимость синхронизации работы отдельных компонентов модели;
* наличие модели рабочей нагрузки, качество которой зависит, в свою очередь, от тех же факторов.

Пригодность имитационной модели для решения задач исследования характеризуется тем, в какой степени она обладает так называемыми целевыми свойствами. Основными из них являются:

* адекватность;
* устойчивость;
* чувствительность.

Ниже рассмотрены некоторые способы проведения оценки модели по каждому из них.

**10.1.1. Оценка адекватности модели.**

В обще случае под адекватностью понимают степень соответствия модели тому реальному явлению или объекту, для описания которого она строится.

Вместе с тем, создаваемая модель ориентирована, как правило, на исследование определенного подмножества свойств этого объекта. Поэтому можно считать, что адекватность модели определяется степенью ее соответствия не столько реальному объекту, сколько целям исследования. В наибольшей степени это утверждение справедливо относительно моделей проектируемых систем (т.е. в ситуациях, когда реальная система вообще не существует).

Тем не менее, во многих случаях полезно иметь формальное подтверждение (или обоснование) адекватности разработанной модели. Один из наиболее распространенных способов такого обоснования - использование методов математической статистики. Суть этих методов заключается в проверке выдвинутой гипотезы (в данном случае - об адекватности модели) на основе некоторых статистических критериев.

*Замечание: при проверке гипотез методами математической статистики необ­ходимо иметь в виду, что статистические критерии не могут доказать ни одной ги­потезы: они могут лишь указать на отсутствие опровержения.*

Процедура оценки основана на сравнении измерений на реальной системе и результатов экспериментов на модели и может проводиться различными способами. Наиболее распространенные из них:

* по средним значениям откликов модели и системы;
* по дисперсиям отклонений откликов модели от среднего значения откликов системы;
* по максимальному значению относительных отклонений откликов  
  модели от откликов системы.

Названные способы оценки достаточно близки по сути, поэтому ограничимся рассмотрением первого из них.

При этом способе проверяется гипотеза о близости среднего значения наблюдаемой переменной У среднему значению отклика реальной системы Y\*.

В результате No опытов на реальной системе получают выборку Y\*. Выполнив NM экспериментов на модели, также получают множество значений наблюдаемой переменной Y.

Затем вычисляются оценки математического ожидания и дисперсии откликов модели и системы, после чего выдвигается гипотеза о близости средних значений величин У\* и У (в статистическом смысле). Основой для проверки гипотезы является t-статистика (распределение Стьюдента). Ее значение, вычисленное по результатам испытаний, сравнивается с критическим значением £кр, взятым из справочной таблицы. Если выполняется неравенство t < £кр, то гипотеза принимается.

*Необходимо еще раз подчеркнуть, что статистические методы применимы только в том случае, если оценивается адекватность модели существующей системе. На проектируемой системе провести измерения, естественно, не представляется возможным. Единственный способ преодолеть это препятствие заключается в том, чтобы принять в качестве эталонного объекта концептуальную модель проектируемой системы. Тогда оценка адекватности программно реализованной модели заключается в проверке того, насколько корректно она отражает концептуальную модель. Данная проблема сходна с проверкой корректности любой компьютерной программы, и ее можно решать соответствующими методами, например с помощью тестирования.*

**10.1.2. Оценка устойчивости модели.**

При оценке адекватности модели как существующей, так и проектируемой системе реально может быть использовано лишь ограниченное подмножество всех возможных значений входных параметров (рабочей нагрузки и внешней среды). В связи с этим для обоснования достоверности получаемых результатов моделирования большое значение имеет проверка устойчивости модели. В теории моделирования это понятие трактуется следующим образом.

*Устойчивость модели* — это ее способность сохранять адекватность при исследовании эффективности системы на всем возможном диапазоне рабочей нагрузки, а также при внесении изменений в конфигурацию системы.

Универсальной процедуры проверки устойчивости модели не существует. Разработчик вынужден прибегать к частичным тестам и здравому смыслу. Часто бывает полезна апостериорная проверка. Она состоит в сравнении результатов моделирования и результатов измерений на системе после внесения в нее изменений. Если результаты моделирования приемлемы, уверенность в устойчивости модели возрастает.

В общем случае можно утверждать, что чем ближе структура модели структуре системы и чем выше степень детализации, тем устойчивее модель.

Устойчивость результатов моделирования может быть также оценена методами математической статистики. Здесь уместно вспомнить основную задачу математической статистики. В данном случае устойчивость результатов моделирования можно рассматривать как признак, подлежащий оценке. Для проверки гипотезы об устойчивости результатов может быть использован критерий Уилкоксона.

Критерий Уилкоксона служит для проверки того, относятся ли две выборки к одной и той же генеральной совокупности (т. е. обладают ли они одним и тем же статистическим признаком). При статистической оценке устойчивости модели соответствующая гипотеза может быть сформулирована следующим образом: при изменении входной (рабочей) нагрузки или структуры ИМ закон распределения результатов моделирования остается неизменным.

Определение однородности двух выборок проводится на основе проверки статистической гипотезы *Н0* о том, что выборки  и принадлежат одному, пусть и неизвестному, закону распределения. При этом применяют критерии знаков, критерий Вилкоксона (Вилкоксона – Мана – Уитни) и другие. Существенной особенностью критерия знаков является требование равного объема сравниваемых выборок, правда, это требование легко выполнить, "обрезав" при рассмотрении более длинную выборку (но такая операция приводит к потере части информации при сопоставлении выборок). В интересах решения поставленной задачи целесообразно применять второй критерий – *критерий Вилкоксона*.

Проверка однородности выборок по критерию Вилкоксона состоит в следующем. Пусть для случайной величины Х имеется выборка объема nx и для случайной величины Y выборка объема ny. По этим выборкам необходимо с уровнем значимости a проверить гипотезу Н0 о том, что функция распределения F(x) случайной величин Х равна функции распределения F(y) случайной величины Y. Конкурирующая гипотеза – функции распределения случайных величин различны: F(x) < F(y) или F(x) > F(y), т.е. критическая область двусторонняя.

Сущность проверки основана на простой идее: если верна гипотеза Н0, то нельзя ожидать преобладания наблюдений одной из выборок на любом из концов вариационного ряда, иначе говоря, результаты наблюдений из каждого слоя должны быть рассеяны по всему вариационному ряду. Такая проверка осуществляется только по порядковым соотношениям x > y и x < y между элементами выборок.

Далее считается, что объем первой выборки не превышает объема второй. Если это условие не выполняется, то выборки просто меняются местами. Проверка гипотезы однородности имеет свою специфику для разных объемов выборок.

Пусть nx >3, ny >3 и суммарный объем обеих выборок не превосходит *25*. Проверка гипотезы осуществляется поэтапно:

* из выборок исключаются одинаковые элементы (вероятность совпадения элементов весьма невелика, поэтому число исключаемых членов выборок не будет большим);
* на основе элементов обеих выборок строится общий вариационный ряд, индексы и конкретные значения элементов можно опустить. В результате получится просто последовательность букв *y* и *x*, например *x x x y x y y x x x y y y* ;
* подсчитывается сумма порядковых номеров *u* вариант первой (меньшей по объему) выборки. В приведенном примере *nx* > *n*y (*nx* = 7 и *ny* = 6), поэтому первой будем считать выборку для величины *Y*. Буква *y* встречается на четвертом, шестом, седьмом, одиннадцатом, двенадцатом и тринадцатом местах, следовательно

*u*=4+6+7+11+12+13=53.

Случайная величина *u* имеет распределение Вилкоксона. Для нее построена специальная таблица нижних критических точек распределения.

* по таблице критических точек для *ny* = 6, *nx* = 7, заданного уровня значимости, например *a* = 0,05 (критическая область двусторонняя, следовательно, каждая сторона критической области соответствует уровню значимости *a*/2 = 0,025), определяется нижняя критическая точка uн. В данном случае *uн* = 27;
* вычисляется верхняя критическая точка *uв* = (*ny+nx+*1)*ny* – *uн*. Для рассматриваемого примера

*uв* = (6+7+1)6 – 27 = 57;

* если *u* < *uн* или *u* > *uв*, то нулевую гипотезу отвергают. В противном случае нет оснований для отклонения нулевой гипотезы. В приведенном примере нулевая гипотеза об однородности выборок принимается.

Сумма порядковых номеров вариант первой выборки с увеличением общего объема выборок стремится к нормальному распределению. *Нормальное распределение можно применять, если nx >3, ny >3 и объем хотя бы одной из выборок превосходит 25*. В таком случае значение нижней критической точки величины *u* при *nx · ny*



где *t1– a/2* – квантиль уровня 1–*a*/2 стандартизованной нормальной случайной величины.

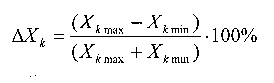
Остальные этапы проверки ничем не отличаются от рассмотренных выше, применительно к малому объему слоев.

**10.1.3. Оценка чувствительности ИМ.**

Очевидно, что устойчивость является положительным свойством модели. Однако если изменение входных воздействий или параметров модели (в некотором заданном диапазоне) не отражается на значениях выходных параметров, то польза от такой модели невелика (ее можно назвать «бесчувственной»). В связи с этим возникает задача оценивания чувствительности модели к изменению параметров рабочей нагрузки и внутренних параметров самой системы.

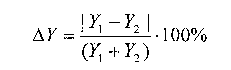
Такую оценку проводят по каждому параметру Хk в отдельности. Основана она на том, что обычно диапазон возможных изменений параметра известен. Одна из наиболее простых и распространенных процедур оценивания состоит в следующем.

1) вычисляется величина относительного среднего приращения параметра Хk:

****

2) проводится пара модельных экспериментов при значениях Хk =Хkmax и Хk = Xkmin и средних фиксированных значениях остальных параметров. Определяются значения отклика модели Y1=f(Xkmax) и Y2=(Xkmin);

3)вычисляется ее относительное приращение наблюдаемой переменной У:



В результате для k-го параметра модели имеют пару значений (∆Хк, ∆Ук), характеризующую чувствительность модели по этому параметру.

Аналогично формируются пары для остальных параметров модели, которые образуют множество {∆Хк,∆Ук}.

Данные, полученные при оценке чувствительности модели, могут быть использованы, в частности, при планировании экспериментов: большее внимание должно уделяться тем параметрам, по которым модель является более чувствительной.

**10.1.4. Калибровка модели.**

Если в результате проведенной оценки качества модели оказалось, что ее целевые свойства не удовлетворяют разработчика, необходимо выполнить ее калибровку, т. е. коррекцию с целью приведения в соответствие предъявляемым требованиям.

Как правило, процесс калибровки носит итеративный характер и состоит из трех основных этапов:

* глобальные изменения модели (например, введение новых процессов, изменение типов событий и т. д.);
* локальные изменения (в частности, изменение некоторых законов  
  распределения моделируемых случайных величин);
* изменение специальных параметров, называемых калибровочными.

Целесообразно объединить оценку целевых свойств ИМ и ее калибровку в единый процесс. Именно такая стратегия принята в статистическом методе калибровки, описанном ниже.

Процедура калибровки состоит из трех шагов, каждый из которых является итеративным (рис. 2.13).

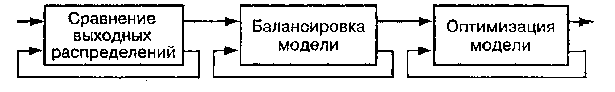


Рис 2.13. Схема процесса калибровки ИМ

**Шаг** 1. Сравнение выходных распределений.

Цель - оценка адекватности ИМ. Критерии сравнения могут быть различны. В частности, может использоваться величина разности между средними значениями откликов модели и системы.

Устранение различий на этом шаге основано на внесении глобальных изменений.

**Шаг** 2. Балансировка модели.

Основная задача - оценка устойчивости и чувствительности модели. По его результатам, как правило, производятся локальные изменения (но возможны и глобальные).

Шаг 3. **Оптимизация модели.**

Цель этого этапа — обеспечение требуемой точности результатов. Здесь возможны три основных направления работ:

* дополнительная проверка качества датчиков СЧ;
* снижение влияния переходного режима;
* применение специальных методов понижения дисперсии.

**10.2. Подбор параметров распределений**

В некоторых случаях имитационная модель сложной системы может быть реализована в виде набора отдельных моделей ее подсистем. При проведении экспериментов с такой моделью в целях сокращения затрат времени бывает необходимо заменять моделирование работы одной из подсистем некоторым числовым параметром (вспомните принцип параметризации), либо случайной величиной, распределенной по заданному закону. Чтобы такая замена была выполнена корректно, исследователь должен располагать описанием зависимости данного числового параметра от времени и других факторов, фигурирующих в модели.

При имитационном моделировании подбор законов распределений выполняется на основе статистических данных, полученных в ходе эксперимента.

В основе процедуры отыскания закона распределения некоторой величины по экспериментальным данным лежит проверка статистических гипотез.

Статистическая гипотеза — это утверждение относительно значений одного или боле параметров распределения некоторой величины или о самой форме распределения.

Обычно выбирают две исходные гипотезы: основную—Н0 и альтернативную ей – H1

Статистическая проверка гипотезы - это процедура выяснения, следует ли принял основную гипотезу Н0 или отвергнуть ее.

Если в результате проверки гипотеза Н0 ошибочно отвергается, то имеет место ошибка 1-го рода (характеризующаяся более тяжелыми последствиями); если гипотеза Н0 принимается при истинности H1 — это ошибка второго рода.

Вероятности ошибок I и II рода (α и β) зависят от критерия, на основание которого будет выбираться одна из гипотез. Очевидно, что вероятности этих двух ошибок взаимосвязаны, то есть чем больше значение α, тем меньше βи наоборот.

Обычное решение этой дилеммы состоит в том, что выбирают некоторое фиксированное значение α(как правило 0.05, 0.01, 0.001) и надеются, что β будет также мало. Фиксированное значение а называется **уровнем значимости.**

Для выбранного значения α определяется так называемая критическая область В, удовлетворяющая условию

****

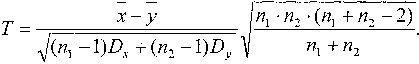
где Z— контрольная величина (критерий), представляющая собой некоторую функцию от выборки (результатов эксперимента).

Проверка гипотезы состоит в следующем. Производится выборка (проводится эксперимент), на основании чего вычисляется z — частное значение критерия Z. Если , то от гипотезы Н0 отказываются. Если z не лежит в В, то говорят, что полученные наблюдения **не противоречат принятой гипотезе.**

Для наиболее часто используемых статистических гипотез разработаны критерии, позволяющие проводить их проверку с наибольшей достоверностью. Рассмотрим основные из них.

***t-критерий***служит для проверки гипотезы о равенстве средних значений двух нормально распределенных СВ (X и Y) в предположении, что дисперсии их равны (хотя и неизвестны). Сравниваемые выборки могут иметь разный объем.

В качестве критерия используют величину

****

Величина Т подчиняется t-распределению Стьюдента.

Критическое значение для t-критерия определяется по таблице для выбранного значения α и числа степеней свободы k = n1+n2-2.

Если вычисленное значение Т превышает tкр, найденное по таблице, то ги­потезу Н0 отвергают.

По отношению к предположению о «нормальной распределенности» величин x и у t-критерий не очень чувствителен. Его можно применять, если распределения СВ не имеют нескольких вершин и не слишком асимметричны.

***F-критерий***служит для проверки гипотезы о равенстве дисперсий Dx и Dy при условии, что x и у распределены нормально.

Гипотезы такого рода имеют большое значение в технике, так как дисперсия есть мера таких характеристик, как погрешности измерительных приборов, точность технологических процессов, точность наведения при стрельбе и так далее.

В качестве контрольной величины используется отношение дисперсий F= Dx / Dy (или Dy / Dx - большая дисперсия должна быть в числителе).

Величина F подчиняется F-распределению (Фишера) с (m1, m2) степенями свободы (m1=n1-1, m2=n2 - l). Проверка гипотезы состоит в следующем.

Для величины γ = α/2 и величин m1 m2 по таблице F-распределения выбирают значения Fγ,m1,m2. Если статистика F\*, вычисленная по выборке, больше этого критического значения, гипотеза должна быть отклонена с вероятностью ошибки α.

Критерии согласия — это критерии, с помощью которых проверяют, удовлетворяет ли рассматриваемая СВ данному закону распределения.

*Критерий согласия Пирсона* **(формула)**и *Критерий Колмогорова* — *Смирнова* были подробно рассмотрены в курсе ТВиМС.

**10.3.Оценка влияния и взаимосвязи факторов**

Как правило, количественная оценка степени влияния того или иного фактора на значения наблюдаемой переменной (показателя эффективности) вызывает значительную сложность, особенно при наличии взаимного влияния факторов. Наиболее простой и доступный способ решения этой проблемы состоит в использовании результатов оценки чувствительности модели.

Однако эти результаты сложно представить в форме аналитической зависимости. Такое представление может оказаться весьма полезным для многих практических задач, связанных как с разработкой моделей (речь опять-таки идет о принципе параметризации), так и непосредственно с принятием решений по экспериментальным данным.

Отыскание аналитических зависимостей, связывающих между собой различные параметры, фигурирующие в модели, может быть основано на совместном использовании группы методов математической статистики: дисперсионного, корреляционного и регрессионного анализа. Подробному и строгому описанию соответствующих процедур посвящено огромное количество книг учебного, научного и справочного характера. Поэтому основная цель изложения последующего материала сводится к тому, чтобы показать роль и место указанных методов при проведении анализа данных, полученных в ходе имитационного эксперимента.

**10.3.1. Однофакторный дисперсионный анализ**

Его суть сводится к определению влияния на результат моделирования одного выбранного фактора.

Пусть, например, исследователя интересует средняя интенсивность отказов компьютера, и в созданной им модели учтены следующие факторы: интенсивность поступления заданий пользователей, интенсивность обращений в оперативную память, временные характеристики решаемых задач и интенсивность обращений к жесткому диску. Если предварительные данные говорят о том, что основной причиной отказов является ненадежная работа жесткого диска, то в качестве анализируемого фактора целесообразно выбрать интенсивность обращений к нему. Задача факторного анализа в данном случае состоит в том, чтобы оценить влияние указанного фактора на среднее число отказов.

*Однофакторный дисперсионный анализ* позволяет установить, оказывает ли существенное влияние некоторый фактор Φ, который имеет несколько уровней, на исследуемую случайную величину.

*Задача сравнения выборок случайных величин* формулируется следующим образом.

Имеются результаты наблюдений в виде совокупности слоев типа

,

задан уровень значимости *a* для проверки статистической гипотезы. В данном случае отдельные слои трактуются как выборки одной и той же случайной величины, полученные по результатам наблюдения за одним объектом при различных значениях фактора Φ (количество уровней фактора равно *m*).

Требуется проверить нулевую гипотезу *Н*0 о равенстве математических ожиданий случайных величин всех выборок. Иначе говоря, требуется установить, значимо или незначимо различаются выборочные средние значения, вычисленные для каждого слоя.

***Допущения:*** генеральные совокупности, соответствующие каждому слою, распределены *нормально*; *дисперсии* слоев *одинаковы*; математические ожидания, дисперсии, законы распределения случайных величин для различных слоев неизвестны, сами случайные величины являются непрерывными. Вполне понятно, что первые два условия являются наиболее существенными и весьма ограничивают область применения методов дисперсионного анализа.

Основная идея дисперсионного анализа состоит не в сопоставлении математических ожиданий случайных величин, а в сравнении оценки "*факторной* *дисперсии*", порождаемой воздействием фактора, и оценки "*остаточной* *дисперсии*", обусловленной случайными причинами. Если различие между этими оценками значимо, то фактор оказывает существенное влияние на случайную величину, в противном случае влияние фактора несущественно. Если установлено существенное влияние фактора, то каждому слою соответствует своя оценка математического ожидания. Упорядочение значений оценок математического ожидания позволит выявить влияние фактора.

Эту же задачу можно было бы решить путем проверки нулевой гипотезы о равенстве минимального и максимального значений оценок математического ожидания, вычисленных по всем слоям. Но такое сопоставление выборок игнорирует информацию, содержащуюся во всех слоях, кроме выбранных, и поэтому нецелесообразно.

Дисперсионный анализ выполняется поэтапно. Такими этапами являются следующие:

* проверка выборок на принадлежность к нормальному закону распределения. Этап необходим, когда нет априорной информации о законах распределения слоев. Сущность такой проверки была рассмотрена в разделе 2 настоящего курса лекций. Если принадлежность нормальному закону не подтвердится, то аппарат дисперсионного анализа, вообще говоря, применять нельзя. Некоторые исследователи допускают его применение при больших объемах выборок (объем каждой выборки должен быть не менее 30) независимо от вида закона распределения;
* проверка равенства оценок дисперсий во всех слоях выборки (проверка однородности дисперсий). Если однородность не подтвердится, то методы дисперсионного анализа не применимы;
* вычисление оценки факторной и остаточной дисперсии;
* сравнение средних значений величин методом дисперсионного анализа и формирование выводов по результатам сравнения.

#### 10.3.1.1 Проверка однородности совокупности дисперсий

Для каждого слоя вычисляется несмещенная оценка дисперсии, обозначим эти оценки через *μ\*2(x)* , *μ\*2* *(*y), …, *μ\*2* *(w)* соответственно. Числа степеней свободы этих оценок

*k1 = п1–*1*, k2 = п2 –* 1*, …, kт = пw–1.*

Гипотеза *Н0* состоит в том, что выборки, по которым определены оценки дисперсии, получены из генеральных совокупностей, обладающих одинаковыми дисперсиями

*μ2(x)= μ2(y)= … =μ2(w)= μ2,*

при этом величина дисперсии *μ2* остается неизвестной. Следует выяснить, являются ли величины *μ\*2(x)* , *μ\*2* *(*y), …, *μ\*2* *(w)* оценками одной и той же генеральной дисперсии μ2.

Рассмотрим сначала случай, когда объем выборок по слоям хотя бы частично различается. В такой ситуации применяется критерий однородности Бартлетта. Проверка однородности реализуется в несколько шагов.

Вычисляется усредненная оценка несмещенной дисперсии по всем слоям

  (10.4)

где μ\*2 (i) – несмещенная оценка дисперсии для слоя i.

Рассчитывается значение критерия

 (10.5)

Бартлетт установил, что случайная величина *В* при условии справедливости нулевой гипотезы распределена приближенно как хи-квадрат с *т*–1 степенями свободы, если все *ni* больше трех. По заданному уровню значимости *a*, числу степеней свободы *т*–1 для правосторонней критической области определяется критическое значение с2кр(m-1;a). Если соблюдается условие

*B< с2кр(m-1;a),*

то нет оснований отвергнуть нулевую гипотезу. Если B > с2кр(m-1;a), то нулевая гипотеза отвергается. Критерий Бартлетта чувствителен к отклонениям распределения от нормального, поэтому к результатам сравнения следует относится осторожно, а при одинаковом объеме всех слоев вместо критерия Бартлетта лучше применять критерий Кочрена (Кохрена).

Итак, если *k1 = k2 = … = kт,* то применяется *критерий Кочрена*

 (10.6)

где *μ*2, max – максимальная оценка дисперсии по всем слоям.

Критическая область для критерия Кочрена правосторонняя. Критическую точку   Gкр(k1,m;α) находят по таблице распределения Кочрена, фрагмент которой приведен в табл. Критическая область определяется неравенством G > Gкр(k1,m;α) .

#### 10.3.1.2. Сравнение факторной и остаточной дисперсий

Пусть все выборки (6.1) характеризуют одну случайную величину Х при различных значениях фактора Φ, т.е. каждый слой соответствует одному количественному или качественному значению фактора. Сравнение дисперсий производится в следующем порядке:

* рассчитывается среднее значение (оценка математического ожидания) по всей совокупности наблюдений

 ,

где *n*=*n*1+*n*2+…+*n*m , а *х*ij *– j*-й *элемент i*-го слоя;

* вычисляются средние значения для всех слоев (групп) ;
* определяется общая сумма квадратов отклонений наблюдаемых значений от оценки математического ожидания

; (10.7)

* определяется факторная сумма квадратов отклонений средних по слоям от оценки математического ожидания (характеризует рассеяние между слоями)

; (10.8)

* определяется остаточная сумма квадратов отклонений наблюдаемых значений внутри слоя от своей средней

. (10.9)

Величина *Sфакт* характеризует влияние фактора Φ. Это положение можно пояснить следующим образом. Пусть фактор оказывает существенное влияние на величину *Х*. Тогда результаты наблюдения для одного слоя, вообще говоря, отличаются от результатов, представленных в других слоях. Следовательно, различаются и средние значения по слоям, причем они тем больше отличаются от оценки математического ожидания по всей выборке, чем больше проявляется влияние фактора. Таким образом, сумма квадратов отклонений средних по слоям от общей средней и характеризует влияние фактора (возведение отклонений во вторую степень исключает взаимную компенсацию положительных и отрицательных отклонений).

Наблюдения внутри одного слоя различаются из-за воздействия случайных причин. Именно сумма квадратов отклонений наблюдаемых значений в каждом слое от среднего значения в слое и характеризует воздействие этих причин, т.е. величина *Sост* отражает суммарное влияние случайных причин на значение величины *Х*.

Величина *Sобщ*, как сумма квадратов отклонений конкретных значений от среднего значения, характеризует суммарное влияние фактора и случайных причин. Можно показать, что

*Sобщ = Sост + Sфакт*,

тогда для вычисления остаточной суммы квадратов можно воспользоваться более простым соотношением

Sост = Sобщ – Sфакт.

Разделив суммы квадратов отклонений на соответствующее число степеней свободы, получим оценки общей, факторной и остаточной дисперсий:

   (10.10)

Если средние значения случайной величины, вычисленные по отдельным выборкам одинаковы, то оценки факторной и остаточной дисперсий являются несмещенными оценками генеральной дисперсии и различаются несущественно. Тогда сопоставление оценок этих дисперсий по критерию Р. Фишера

*F = μ2факт /μ2ост*

должно показать, что нулевую гипотезу о равенстве факторной и остаточной дисперсий отвергнуть нет оснований. Если *μ2факт < μ2ост*,  то нет необходимости прибегать к вычислению критерия Р. Фишера – из неравенства сразу следует вывод о выполнении нулевой гипотезы. Итак, из справедливости гипотезы о равенстве средних величин по группам следует соблюдение гипотезы о равенстве факторной и остаточной дисперсий.

Если нулевая гипотеза о равенстве средних величин по слоям является ложной, то с увеличением расхождения между слоями возрастает оценка факторной дисперсии, а вместе с ней и величина критерия *F = μ2факт /μ2ост*. В результате значение *F* превысит критическое значение, и гипотеза о равенстве дисперсий будет отвергнута.

Рассуждая от противного, можно доказать справедливость утверждений: из справедливости (ложности) гипотезы о дисперсиях следует истинность (ложность) гипотезы о математических ожиданиях. Таким образом, вместо проверки нулевой гипотезы *Н0* о равенстве средних значений для совокупности выборок следует проверить гипотезу о равенстве факторной и остаточной дисперсий.

*Пример 6.2.* Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве средних значений по слоям, применительно к результатам наблюдений

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Уровень  фактора | Выходная характеристика | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 259,14 | 260,06 | 260,97 | 262.43 | 267.83 | 273.14 |  |
| 2 | 253.68 | 258.14 | 259.49 | 260.18 | 263.65 | 271.39 | 274.12 |
| 3 | 256.36 | 259.36 | 262.84 | 265.94 | 270.33 | 270.44 | 271.63 |

Предполагается, что выборки принадлежат нормальному распределению, а каждый слой соответствует некоторому значению фактора Φ.

*Решение.* Необходимо проверить однородность дисперсий, а затем непосредственно провести дисперсионный анализ. Проверим гипотезу об однородности дисперсий. Для этого вычислим:

* оценки математического ожидания по слоям (групповые средние)

*mгр 1* = 263,93; *mгр 2* = 262,95; *mгр 3*= 265,32;

* несмещенные оценки дисперсии по слоям

*m\*2 (1) = 29,79; m\*2 (2) = 54,20; m\*2 (3) = 34,61;*

* усредненную оценку несмещенной дисперсии по всем слоям

*m\*2* =(29,79×5+54,20×6+34,61×6)/17 = 40,11.

* значение критерия Бартлетта

*В*=*a/c*=0,56/1,08=0,52,

где  *а*=2,303(17lg 40,11 – (5 lg 29,79+6 lg 54,20+6 lg 34,61)) = 0,56;

*с* = 1+(1/5+1/6+1/6 –1/17)/[3(3–1)]=1,08.

Критическое значение хи-квадрат для правосторонней области

*c2кр(*2; 0,05)=6,0.

Поскольку величина *В* меньше *c2кр*(2; 0,05), отвергнуть нулевую гипотезу об однородности дисперсий нет оснований.

Дисперсионный анализ предусматривает вычисление:

* суммы квадратов

*Sобщ* = 701,65; *Sфакт* = 19,81; *Sост* = 681,84;

* оценок дисперсий

*μ\*2* общ=701,65/19=36,93; *μ\*2*  факт=19,81/2=9,91; *μ\*2* ост=681,84/17=40,10.

Оценка факторной дисперсии меньше оценки остаточной дисперсии, поэтому можно сразу утверждать справедливость нулевой гипотезы о равенстве математических ожиданий по слоям выборки, Иначе говоря, в данном примере фактор Φ не оказывает существенного влияния на случайную величину.

**10.3.2. Многофакторный дисперсионный анализ**

Многофакторный дисперсионный анализ (МДА) позволяет оценивать влияние на наблюдаемую переменную уже не одного, а произвольного числа факторов. Точнее, МДА позволяет выбрать ив группы факторов, участвующих в эксперименте, те, которые действительно влияют на его результат.

Методику проведения многофакторного дисперсионного анализа рассмотрим применительно к частичному факторному эксперименту, проводимому в соответствии с латинским планом.

Пусть в эксперименте рассматриваются один первичный фактор и два вторичных, каждый из которых имеет n уровней (т. е. объем испытаний N = n2).

Обозначим через yijk результат эксперимента при условии, что фактор ***а***находился на уровне i, фактор ***b***- на уровне j, фактор ***с***- на уровне к. Множество значений, которые может принимать упорядоченная тройка (i, j, к), обозначим через L.

В этом случае уравнение дисперсионного анализа выглядит следующим образом:

*yijk =m+ai+bj+gk+eijk,*

где m - генеральное среднее случайной величины у,

ai, bj, gk - неизвестные параметры («эффекты» соответствующих факторов).

Решение задачи дисперсионного анализа заключается в проверке гипотез о независимости результатов измерений от факторов ***a, b, c****:*

*Ha: ai=0, i=1, …, n;*

*Hb: bj=0, j=1, …, n;*

*Hc: ck=0, k=1,…, n;*

Для этого по методу наименьших квадратов (МНК) находят оценки параметров m, ai, bj, gk, минимизируя по указанным переменным (поочередно) функцию



Затем по каждому фактору вычисляется f-статистика. Величина F есть мера потерь при принятии гипотезы H0. Чем больше F, тем хуже модель, отвергающая влияние соответствующего фактора. Таким образом, если вычисленное значение F больше Fкр, найденного по таблице для некоторого уровня значимости, то гипотеза отвергается.

Необходимо отметить, что дисперсионный анализ может использоваться для оценки влияния факторов, имеющих как количественный характер, так и качественный, поскольку в уравнении дисперсионного анализа фигурируют не сами факторы, а только их «эффекты».

В том случае, если все факторы носят количественный характер, взаимосвязь между ними и наблюдаемой переменной может быть описана с помощью уравнения регрессии.

**10.3.3. Корреляционный и регрессионный анализ.**

Это два близких метода, которые обычно используются совместно для исследования взаимосвязи между двумя или более непрерывными переменными.

Методы корреляционного анализа позволяют делать статистические выводы о степени зависимости между переменными.

Величина линейной зависимости между двумя переменными измеряется посредством простого коэффициента корреляции, величина зависимости от нескольких - посредством множественного коэффициента корреляции.

В корреляционном анализе используется также понятие частного коэффициента корреляции, который измеряет линейную взаимосвязь между двумя переменными без учета влияния других переменных.

Если корреляционный анализ позволил установить наличие линейной зависимости наблюдаемой переменной от одной или более независимых, то форма зависимости может быть уточнена методами регрессионного анализа.

Для этого строится так называемое уравнение регрессии, которое связывает зависимую переменную с независимыми и содержит неизвестные параметры. Если уравнение линейно относительно параметров (но необязательно линейно относительно независимых переменных), то говорят о линейной регрессии, в противном случае регрессия нелинейна.

#### 10.3.3.1 МАТРИЦА ДАННЫХ

Многие объекты исследования характеризуются множеством параметров, и по результатам наблюдения за их функционированием формируются многомерные совокупности (матрицы) ЭД



Строки такой матрицы соответствуют результатам регистрации всех наблюдаемых параметров объекта в одном эксперименте, а столбцы содержат результаты наблюдений за одним параметром (фактором, вариантой) во всех экспериментах. Обозначим количество параметров через m (m>1), а количество наблюдений – через *n*.

В матрице элемент *хij* соответствует значению *j*-й варианты в *i*-м наблюдении. Матрица, вообще говоря, может содержать пустые значения некоторых элементов, например, из-за пропусков в регистрации значений параметров. В многомерном анализе желательно устранить пропущенные значения. Для этого существуют специальные приемы, в частности, вычеркивание соответствующих строк матрицы или занесение средних значений вместо отсутствующих. В дальнейшем будем считать, что матрица не содержит пустых элементов, а параметры объекта характеризуются непрерывными случайными величинами.

Методы обработки матрицы ЭД основаны на следующем предположении: если объект подвергнуть новому обследованию и получить, вообще говоря, другую матрицу данных, то после ее обработки с помощью тех же методов будут получены результаты, близкие к результатам обработки первой матрицы. Данное предположение основано на статистической гипотезе формирования матрицы ЭД. Матрица порождается случайным образом в соответствии с определенной вероятностной закономерностью, а именно: в *m*-мерном пространстве параметров существует некоторое (пусть и неизвестное) распределение вероятностей, и каждая строка матрицы появляется в соответствии с этим распределением независимо от появления других строк.

Каждый столбец матрицы представляет собой случайную выборку значений одного параметра объекта. Указанное предположение означает, во-первых, что оценки моментов и параметров распределения, вычисленные по выборке, будут близки к истинным значениям, во-вторых, значения непрерывных функций, построенных по этим оценкам, будут близки к значениям функций, построенным по истинным значениям параметров.

Таким образом, объектом исследования в многомерном анализе является многомерная случайная величина, представленная выборкой конечного объема. К такой выборке применимы все методы и оценки, рассмотренные при обработке одномерных ЭД. Конечно, приведенные суждения не являются доказательством допустимости применения рассматриваемых методов, но вполне подтверждаются практикой.

Параметры, характеризующие объект исследования, имеют разный физический смысл, и матрица данных существенно изменяется, если изменяются шкалы, в которых измеряются те или иные параметры. Матрицу данных еще до проведения анализа целесообразно привести к стандартному виду, т.е. стандартизовать значения вариант (напомним, что среднее значение стандартизованной варианты равно нулю, дисперсия – единице). В тех случаях, когда все варианты измеряются в одной шкале, это преобразование все-таки желательно, ибо оно упрощает последующие преобразования. Стандартизованную матрицу будем обозначать через ***U***. Переход от исходной к стандартизованной матрице осуществляется следующим образом.

1. По каждой варианте  вычисляются оценки:

* математического ожидания ;
* дисперсии  .

2. Вычисляются элементы стандартизованной матрицы



Элементы матрицы ***U*** являются безразмерными величинами. Именно матрица ***U*** будет являться объектом последующей обработки.

#### 10.3.3.2. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Величины, характеризующие различные свойства объектов, могут быть независимыми или взаимосвязанными. Различают два вида зависимостей между величинами (факторами): функциональную и статистическую.

При функциональной зависимости двух величин значению одной из них обязательно соответствует одно или несколько точно определенных значений другой величины. Функциональная связь двух факторов возможна лишь при условии, что вторая величина зависит только от первой и не зависит ни от каких других величин. Функциональная связь одной величины с множеством других возможна, если эта величина зависит только от этого множества факторов. В реальных ситуациях существует бесконечно большое количество свойств самого объекта и внешней среды, влияющих друг на друга, поэтому такого рода связи не существуют, иначе говоря, функциональные связи являются математическими абстракциями. Их применение допустимо тогда, когда соответствующая величина в основном зависит от соответствующих факторов.

При исследовании сложных систем многие параметры следует считать случайными, что исключает проявление однозначного соответствия значений. Воздействие общих факторов, наличие объективных закономерностей в поведении объектов приводят лишь к проявлению статистической зависимости. Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения других (другой), и эти другие величины принимают некоторые значения с определенными вероятностями. Функциональную зависимость в таком случае следует считать частным случаем статистической: значению одного фактора соответствуют значения других факторов с вероятностью, равной единице. Однако на практике такое рассмотрение функциональной связи применения не нашло.

Более важным частным случаем статистической зависимости является корреляционная зависимость, характеризующая взаимосвязь значений одних случайных величин со средним значением других, хотя в каждом отдельном случае любая взаимосвязанная величина может принимать различные значения.

Если же у взаимосвязанных величин вариацию имеет только одна переменная, а другая является детерминированной, то такую связь называют не корреляционной, а регрессионной. Например, при анализе скорости обмена с жесткими дисками можно оценивать регрессию этой характеристики на определенные модели, но не следует говорить о корреляции между моделью и скоростью.

При исследовании зависимости между одной величиной и такими характеристиками другой, как, например, моменты старших порядков (а не среднее значение), то эта связь будет называться статистической, а не корреляционной.

Корреляционная связь описывает следующие виды зависимостей:

причинную зависимость между значениями параметров.

"зависимость" между следствиями общей причины.

Корреляционная зависимость определяется различными параметрами, среди которых наибольшее распространение получили показатели, характеризующие взаимосвязь двух случайных величин (парные показатели): корреляционный момент, коэффициент корреляции.

Оценка *корреляционного момента (коэффициента ковариации*) двух вариант *xj* и *xk* вычисляется по исходной матрице ***Х***

 (10.3.2)

Этот показатель неудобен для практического применения, так как имеет размерность, равную произведению размерностей вариант, и по его величине трудно судить о зависимости параметров.

Коэффициент ковариации *rjk* нормированных случайных величин называют коэффициентом корреляции, его оценка

 (10.3.3)

Значение коэффициента корреляции лежит в пределах от –1 до +1. Если случайные величины *Xj* и *Xk* независимы, то коэффициент *rjk* обязательно равен нулю, обратное утверждение неверно. Коэффициент *rjk* характеризует значимость линейной связи между параметрами:

* при *rjk*=1 значения *xij* и *xik* полностью совпадают, т.е. значения параметров принимают одинаковые значения. Иначе говоря, имеет место функциональная зависимость: зная значение одного параметра, можно однозначно указать значение другого параметра;
* при *rjk*= – 1 величины *xij* и *xik* принимают противоположные значения. И в этом случае имеет место функциональная зависимость;
* при *rjk* = 0 величины *xij* и *xik* практически не связаны друг с другом линейным соотношением. Это не означает отсутствия каких-то других (например, нелинейных) связей между параметрами;
* при | *rjk* | > 0 и | *rjk* | < 1 однозначной линейной связи величин *xij* и *xik*нет. И чем меньше абсолютная величина коэффициента корреляции, тем в меньшей степени по значениям одного параметра можно предсказать значение другого.

Используя понятие коэффициента корреляции, матрице ЭД можно поставить в соответствие квадратную матрицу оценок коэффициентов корреляции (корреляционную матрицу)

. (10.3.4)

К числу характерных свойств корреляционной матрицы относят: симметричность относительно главной диагонали, *r\*jk*=*r\*kj*, ; единичные значения элементов главной диагонали, *rkk*=1 (*rkk* соответствует дисперсии стандартизованного параметра *Xk*), .

Оценка коэффициента корреляции, вычисленная по ограниченной выборке, практически всегда отличается от нуля. Но из этого еще не следует, что коэффициент корреляции генеральной совокупности также отличен от нуля. Требуется оценить значимость выборочной величины коэффициента или, в соответствии с постановкой задач проверки статистических гипотез, проверить гипотезу о равенстве нулю коэффициента корреляции. Если гипотеза *Н0* о равенстве нулю коэффициента корреляции будет отвергнута, то выборочный коэффициент значим, а соответствующие величины связаны линейным соотношением. Если гипотеза *Н0* будет принята, то оценка коэффициента не значима, и величины линейно не связаны друг с другом (если по физическим соображениям факторы могут быть связаны, то лучше говорить о том, что по имеющимся ЭД эта взаимосвязь не установлена). Проверка гипотезы о значимости оценки коэффициента корреляции требует знания распределения этой случайной величины. Распределение величины *rik* изучено только для частного случая, когда случайные величины *Uj* и *Uk* распределены по нормальному закону.

В качестве критерия проверки нулевой гипотезы *Н0* применяют случайную величину

.

Если модуль коэффициента корреляции относительно далек от единицы, то величина *t* при справедливости нулевой гипотезы распределена по закону Стьюдента с  *n–2* степенями свободы. Конкурирующая гипотеза *Н1* соответствует утверждению, что значение *rik* не равно нулю (больше или меньше нуля). Поэтому критическая область двусторонняя.

Проверка гипотезы *Н0* о равенстве нулю генерального коэффициента парной корреляции двумерной нормально распределенной случайной величины осуществляется в следующей последовательности:

* вычисляется значение статистики *t;*
* при уровне значимости *a* для двусторонней области определяется критическая точка распределения Стьюдента *tкр(n–2; a)*, табл;
* сравнивается значение статистики *t* с критическим значением *tкр(n–2; a)*. Если *t < tкр (п–2; a)*, то нет оснований отвергнуть нулевую гипотезу, иначе гипотеза *Н0* отвергается (коэффициент корреляции значим).

Когда модуль величины *r\*ik* близок к единице, распределение *r\*ik* отличается от распределения Стьюдента, так как значение |*r\*ik* | ограничено справа единицей. В этом случае применяют преобразование

*yik=0,5ln[(1+|rik |)/(1–|rik |)].*

Величина *yik* не имеет указанного ограничения, она при *п* > 10 распределена приблизительно нормально с центром

*m1\*(rik)=*0,5ln*[(*1*+|rik|)/(*1*–|rik|)]+*0,5*|rik|/(n–*1*)*

и дисперсией

*μ2\* (rik)=s2(rik)=*1*/(п–*3*).*

Если значение центрированной и нормированной величины

*(yik – m1\* (rik))/s(rik)*

превышает значение квантили уровня 1–*a*/2 нормального распределения стандартизованной величины, то нулевая гипотеза отвергается.

Таким образом, *постановка задачи линейного корреляционного анализа* формулируется в следующем виде.

Имеется матрица наблюдений вида (10.3.1).

Необходимо определить оценки коэффициентов корреляции для всех или только для заданных пар параметров и оценить их значимость. Незначимые оценки приравниваются к нулю.

Допущения:

* выборка имеет достаточный объем. Понятие достаточного объема зависит от целей анализа, требуемой точности и надежности оценки коэффициентов корреляции, от количества факторов. Минимально допустимым считается объем, когда количество наблюдений не менее чем в 5–6 раз превосходит количество факторов;
* выборки по каждому фактору являются однородными. Это допущение обеспечивает несмещенную оценку средних величин;
* матрица наблюдений не содержит пропусков.

Если необходима проверка значимости оценки коэффициента корреляции, то требуется соблюдение дополнительного условия – распределение вариант должно подчиняться нормальному закону.

Задача анализа решается в несколько этапов:

* проводится стандартизация исходной матрицы;
* вычисляются парные оценки коэффициентов корреляции;
* проверяется значимость оценок коэффициентов корреляции, незначимые оценки приравниваются к нулю. По результатам проверки делается вывод о наличии связей между вариантами (факторами).

*Пример 10.2.* Результаты наблюдений табл. 10.2.

 Таблица 10.1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| №  пп | Выходная характеристика | Ф1 | Ф2 | Ф3 | Ф4 |
| Х1 | Х2 | Х3 | Х4 | Х5 |
| 1 | 26.37 | 41.98 | 17.66 | 16.05 | 22.85 |
| 2 | 28.00 | 43.83 | 17.15 | 15.47 | 23.25 |
| 3 | 27.83 | 42.83 | 15.38 | 17.59 | 24.55 |
| 4 | 31.67 | 47.28 | 18.39 | 16.92 | 26.59 |
| 5 | 23.50 | 38.75 | 18.32 | 15.66 | 26.22 |
| 6 | 21.04 | 35.12 | 17.81 | 17.00 | 27.52 |
| 7 | 16.94 | 32.07 | 21.42 | 16.77 | 25.76 |
| 8 | 37.56 | 54.25 | 26.42 | 15.68 | 23.10 |
| 9 | 18.84 | 32.70 | 17.23 | 15.92 | 23.41 |
| 10 | 25.77 | 40.51 | 30.43 | 15.29 | 25.17 |
| 11 | 33.52 | 49.78 | 21.71 | 15.61 | 25.39 |
| 12 | 28.21 | 43.84 | 28.33 | 15.70 | 24.56 |
| 13 | 28.76 | 44.03 | 30.42 | 16.87 | 24.45 |
| 14 | 24.60 | 39.46 | 21.66 | 15.25 | 23.81 |
| 15 | 24.51 | 38.78 | 25.77 | 16.05 | 24.48 |

Необходимо определить наличие линейных корреляционных связей между пропускной способностью и остальными факторами. Предполагается, что выборки по всем вариантам подчиняются нормальному закону. Проверку гипотезы о значимости оценок коэффициентов корреляции произвести с уровнем значимости *a*, равным 0,1.

*Решение.* Стандартизация исходной матрицы начинается с вычисления выборочной средней *m1\**, несмещенной оценки дисперсии *μ2\** и среднеквадратического отклонения *S* по каждой варианте, табл.7.2.

Таблица 7.2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Оценка параметра распределения | Варианта | | | | |
| Х1 | Х2 | Х3 | Х4 | Х5 |
| M\* | 26.47 | 41.68 | 21.87 | 16.12 | 24.74 |
| D\* | 29.10 | 36.47 | 26.37 | 0.52 | 1.88 |
| S | 5.39 | 6.04 | 5.13 | 0.72 | 1.37 |

В результате перехода к величинам  формируется стандартизованная матрица исходных данных, табл. 10.3.

 Таблица 10.3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| №  пп | Выходная характеристика | Ф1 | Ф2 | Ф3 | Ф4 |
| U1 | U2 | U3 | U4 | U5 |
| 1 | -0.02 | -0.05 | -0.82 | -0.10 | -1.38 |
| 2 | 0.28 | 0.36 | -0.92 | -0.90 | -1.09 |
| 3 | 0.25 | 0.19 | -1.26 | 2.03 | -0.14 |
| 4 | 0.96 | 0.93 | -0.68 | 1.10 | 2.35 |
| 5 | -0.55 | -0.49 | -0.69 | -0.64 | 1.08 |
| 6 | -1.01 | -1.09 | -0.79 | 1.21 | 2.03 |
| 7 | -1.77 | -1.59 | -0.09 | 0.90 | 0.74 |
| 8 | 2.06 | 2.08 | 0.89 | -0.61 | -1.20 |
| 9 | -1.42 | -1.49 | -0.90 | -0.28 | -0.97 |
| 10 | -0.13 | -0.19 | 1.67 | -1.15 | 0.31 |
| 11 | 1.31 | 1.34 | -0.03 | -0.71 | 0.47 |
| 12 | 0.32 | 0.36 | 1.26 | -0.58 | -0.13 |
| 13 | 0.42 | 0.39 | 1.66 | 1.03 | -0.21 |
| 14 | -0.35 | -0.37 | -0.04 | -1.21 | -0.68 |
| 15 | -0.36 | -0.48 | 0.76 | -0.19 | -0.19 |

Оценки коэффициентов корреляции

 (k = 2, 3, 4)

представлены в табл. 7.4. В этой же таблице приведены значения статистик критерия Стьюдента  для вычисленных оценок коэффициентов корреляции при *п* = 15.

Таблица 7.4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Х2 | Х3 | Х4 | Х5 |
| R1j | 0.93 | 0.25 | -0.13 | -0.22 |
| t | 9.12 | 0.93 | 0.47 | 0.81 |

Критическое значение

*tкр (n–2; a) = tкр (13; 0,1) =* 1,77*.*

Статистика критерия больше критического значения только для *r12*. Это означает, что только для указанного коэффициента оценка значима (коэффициент корреляции генеральной совокупности не равен нулю), а остальные коэффициенты следует признать равными нулю.

Корреляционная зависимость не обязательно устанавливается только для двух величин, с ее помощью можно анализировать связи между несколькими вариантами (множественная корреляция). А кроме линейной существуют и другие виды корреляции.

#### 10.3.3.3. РЕГРЕССИОННЫЙ АНАЛИЗ

##### *10.3.3.3.1. Постановка задачи*

Одной из типовых задач обработки многомерных ЭД является определение количественной зависимости показателей качества объекта от значений его параметров и характеристик внешней среды. Примером такой постановки задачи является установление зависимости между временем обработки запросов к базе данных и интенсивностью входного потока. Время обработки зависит от многих факторов, в том числе от размещения искомой информации на внешних носителях, сложности запроса. Следовательно, время обработки конкретного запроса можно считать случайной величиной. Но вместе с тем, при увеличении интенсивности потока запросов следует ожидать возрастания его среднего значения, т.е. считать, что время обработки и интенсивность потока запросов связаны корреляционной зависимостью.

*Постановка задачи регрессионного анализа* формулируется следующим образом.

Имеется совокупность результатов наблюдений вида (7.1). В этой совокупности один столбец соответствует показателю, для которого необходимо установить функциональную зависимость с параметрами объекта и среды, представленными остальными столбцами. Будем обозначать показатель через *y\** и считать, что ему соответствует первый столбец матрицы наблюдений. Остальные *т*–1 (*m* > 1) столбцов соответствуют параметрам (факторам) *х2, х3, …, хт* .

Требуется: установить количественную взаимосвязь между показателем и факторами. В таком случае задача регрессионного анализа понимается как задача выявления такой функциональной зависимости *y\* = f(x2 , x3 , …, xт*), которая наилучшим образом описывает имеющиеся экспериментальные данные.

Допущения:

* количество наблюдений достаточно для проявления статистических закономерностей относительно факторов и их взаимосвязей;
* обрабатываемые ЭД содержат некоторые ошибки (помехи), обусловленные погрешностями измерений, воздействием неучтенных случайных факторов;
* матрица результатов наблюдений является единственной информацией об изучаемом объекте, имеющейся в распоряжении перед началом исследования.

Функция *f(x2 , x3 , …, xт)*, описывающая зависимость показателя от параметров, называется уравнением (функцией) регрессии. Термин "регрессия" (regression (лат.) – отступление, возврат к чему-либо) связан со спецификой одной из конкретных задач, решенных на стадии становления метода. Его ввел английский статистик Ф. Гальтон. Он исследовал влияние роста родителей и более отдаленных предков на рост детей. По его модели рост ребенка определяется наполовину родителями, на четверть – дедом с бабкой, на одну восьмую прадедом и прабабкой и т.д. Другими словами, такая модель характеризует движение назад по генеалогическому дереву. Ф. Гальтон назвал это явление регрессией как противоположное движению вперед – прогрессу. В настоящее время термин "регрессия" применяется в более широком плане – для описания любой статистической связи между случайными величинами.

Решение задачи регрессионного анализа целесообразно разбить на несколько этапов:

* предварительная обработка ЭД;
* выбор вида уравнений регрессии;
* вычисление коэффициентов уравнения регрессии;
* проверка адекватности построенной функции результатам наблюдений.

Предварительная обработка включает стандартизацию матрицы ЭД, расчет коэффициентов корреляции, проверку их значимости и исключение из рассмотрения незначимых параметров (эти преобразования были рассмотрены в рамках корреляционного анализа). В результате преобразований будут получены стандартизованная матрица наблюдений ***U*** (через *y* будем обозначать стандартизованную величину *y\**) и корреляционная матрица **r**.

Стандартизованной матрице ***U*** можно сопоставить одну из следующих геометрических интерпретаций:

* в *т*-мерном пространстве оси соответствуют отдельным параметрам и показателю. Каждая строка матрицы представляет вектор в этом пространстве, а вся матрица – совокупность *п* векторов в пространстве параметров;
* в *п*-мерном пространстве оси соответствуют результатам отдельных наблюдений. Каждый столбец матрицы – вектор в пространстве наблюдений. Все вектора в этом пространстве имеют одинаковую длину, равную . Тогда угол между двумя векторами характеризует взаимосвязь соответствующих величин. И чем меньше угол, тем теснее связь (тем больше коэффициент корреляции).

В корреляционной матрице особую роль играют элементы левого столбца – они характеризуют наличие или отсутствие линейной зависимости между соответствующим параметром *ui* (i =2, 3, …, *т*) и показателем объекта *y*. Проверка значимости позволяет выявить такие параметры, которые следует исключить из рассмотрения при формировании линейной функциональной зависимости, и тем самым упростить последующую обработку.

##### *10.3.3.3.2. Выбор вида уравнения регрессии*

Задача определения функциональной зависимости, наилучшим образом описывающей ЭД, связана с преодолением ряда принципиальных трудностей. В общем случае для стандартизованных данных функциональную зависимость показателя от параметров можно представить в виде

, (10.3.5)

где *f* – заранее не известная функция, подлежащая определению;

ε - ошибка аппроксимации ЭД.

Указанное уравнение принято называть выборочным уравнением регрессии *y* на *u*. Это уравнение характеризует зависимость между вариацией показателя и вариациями факторов. А мера корреляции измеряет долю вариации показателя, которая связана с вариацией факторов. Иначе говоря, корреляцию показателя и факторов нельзя трактовать как связь их уровней, а регрессионный анализ не объясняет роли факторов в создании показателя.

Еще одна особенность касается оценки степени влияния каждого фактора на показатель. Регрессионное уравнение не обеспечивает оценку раздельного влияния каждого фактора на показатель, такая оценка возможна лишь в случае, когда все другие факторы не связаны с изучаемым. Если изучаемый фактор связан с другими, влияющими на показатель, то будет получена смешанная характеристика влияния фактора. Эта характеристика содержит как непосредственное влияние фактора, так и опосредованное влияние, оказанное через связь с другими факторами и их влиянием на показатель.

В регрессионное уравнение не рекомендуется включать факторы, слабо связанные с показателем, но тесно связанные с другими факторами. Не включают в уравнение и факторы, функционально связанные друг с другом (для них коэффициент корреляции равен 1). Включение таких факторов приводит к вырождению системы уравнений для оценок коэффициентов регрессии и к неопределенности решения.

 Функция *f* должна подбираться так, чтобы ошибка ε в некотором смысле была минимальна. Существует бесконечное множество функций, описывающих ЭД абсолютно точно (ε = 0), т.е. таких функций, которые для всех значений параметров *uj,2 , uj,3 , …, uj,т*принимают в точности соответствующие значения показателя *yi , i* =1, 2, …, *п*. Вместе с тем, для всех других значений параметров, отсутствующих в результатах наблюдений, значения показателя могут принимать любые значения. Понятно, что такие функции не соответствуют действительной связи между параметрами и показателем.

В целях выбора функциональной связи заранее выдвигают гипотезу о том, к какому классу может принадлежать функция *f*, а затем подбирают "лучшую" функцию в этом классе. Выбранный класс функций должен обладать некоторой "гладкостью", т.е. "небольшие" изменения значений аргументов должны вызывать "небольшие" изменения значений функции (ЭД содержат некоторые ошибки измерений, а само поведение объекта подвержено влиянию помех, маскирующих истинную связь между параметрами и показателем).

Простым, удобным для практического применения и отвечающим указанному условию является класс полиномиальных функций

 (10.3.6)

Для такого класса задача выбора функции сводится к задаче выбора значений коэффициентов *a0 , aj , ajk , …, ajj , …* . Однако универсальность полиномиального представления обеспечивается только при возможности неограниченного увеличения степени полинома, что не всегда допустимо на практике, поэтому приходится применять и другие виды функций.

Частным случаем, широко применяемым на практике, является полином первой степени или уравнение линейной регрессии

. (7.7)

Это уравнение в регрессионном анализе следует трактовать как векторное, ибо речь идет о матрице данных

, i =1, 2, … ,  *n*. (10.3.8)

Обычно стремятся обеспечить такое количество наблюдений, которое превышало бы количество оцениваемых коэффициентов модели. Для линейной регрессии при *п* *> т* количество уравнений превышает количество подлежащих определению коэффициентов полинома. Но и в этом случае нельзя подобрать коэффициенты таким образом, чтобы ошибка в каждом скалярном уравнении обращалась в ноль, так как к неизвестным относятся *аj* и *ei*, их количество *n* + *т* – 1, т.е. всегда больше количества уравнений *п*. Аналогичные рассуждения справедливы и для полиномов степени, выше первой.

Для выбора вида функциональной зависимости можно рекомендовать следующий подход:

* в пространстве параметров графически отображают точки со значениями показателя. При большом количестве параметров можно строить точки применительно к каждому из них, получая двумерные распределения значений;
* по расположению точек и на основе анализа сущности взаимосвязи показателя и параметров объекта делают заключение о примерном виде регрессии или ее возможных вариантах;
* после расчета параметров оценивают качество аппроксимации, т.е. оценивают степень близости расчетных и фактических значений;
* если расчетные и фактические значения близки во всей области задания, то задачу регрессионного анализа можно считать решенной. В противном случае можно попытаться выбрать другой вид полинома или другую аналитическую функцию, например периодическую.

##### *10.3.3.3.3 Вычисление коэффициентов уравнения регрессии*

Систему уравнений (10.3.8) на основе имеющихся ЭД однозначно решить невозможно, так как количество неизвестных всегда больше количества уравнений. Для преодоления этой проблемы нужны дополнительные допущения. Здравый смысл подсказывает: желательно выбрать коэффициенты полинома так, чтобы обеспечить минимум ошибки аппроксимации ЭД. Могут применяться различные меры для оценки ошибок аппроксимации. В качестве такой меры нашла широкое применение среднеквадратическая ошибка. На ее основе разработан специальный метод оценки коэффициентов уравнений регрессии – метод наименьших квадратов (МНК). Этот метод позволяет получить оценки максимального правдоподобия неизвестных коэффициентов уравнения регрессии при нормальном распределения вариант, но его можно применять и при любом другом распределении факторов.

В основе МНК лежат следующие положения:

* значения величин ошибок и факторов независимы, а значит, и некоррелированы, т.е. предполагается, что механизмы порождения помехи не связаны с механизмом формирования значений факторов;
* математическое ожидание ошибки ε должно быть равно нулю (постоянная составляющая входит в коэффициент *a0*), иначе говоря, ошибка является центрированной величиной;
* выборочная оценка дисперсии ошибки  должна быть минимальна.

 Рассмотрим применение МНК применительно к линейной регрессии стандартизованных величин. Для центрированных величин *uj*коэффициент *a0* равен нулю, тогда уравнения линейной регрессии

. (10.3.9)

Здесь введен специальный знак "^", обозначающий значения показателя, рассчитанные по уравнению регрессии, в отличие от значений, полученных по результатам наблюдений.

По МНК определяются такие значения коэффициентов уравнения регрессии, которые обеспечивают безусловный минимум выражению

. (10.3.10)

Минимум находится приравниванием нулю всех частных производных выражения (7.10), взятых по неизвестным коэффициентам, и решением системы уравнений

 (10.3.11)

Последовательно проведя преобразования и используя введенные ранее оценки коэффициентов корреляции



 получим

. (10.3.12)

Итак, получено *т*–1 линейных уравнений, что позволяет однозначно вычислить значения *a2, a3, …, aт*.

Если же линейная модель неточна или параметры измеряются неточно, то и в этом случае МНК позволяет найти такие значения коэффициентов, при которых линейная модель наилучшим образом описывает реальный объект в смысле выбранного критерия среднеквадратического отклонения.

Когда имеется только один параметр, уравнение линейной регрессии примет вид



Коэффициент *a2* находится из уравнения



Тогда, учитывая, что *r 2,2* = 1, искомый коэффициент

*a*2 *= r y*,2. (10.3.13)

Соотношение (7.13) подтверждает ранее высказанное утверждение, что коэффициент корреляции является мерой линейной связи двух стандартизованных параметров.

Подставив найденное значение коэффициента *a2* в выражение для *w*, с учетом свойств центрированных и нормированных величин, получим минимальное значение этой функции, равное 1– *r2y*,2. Величину 1– *r2y,2*называют остаточной дисперсией случайной величины *y* относительно случайной величины *u2*. Она характеризует ошибку, которая получается при замене показателя функцией от параметра υ= *a2u2* . Только при |*ry,2* | = 1 остаточная дисперсия равна нулю, и, следовательно, не возникает ошибки при аппроксимации показателя линейной функцией.

Переходя от центрированных и нормированных значений показателя и параметра



можно получить для исходных величин

 (10.3.14)

 Это уравнение также линейно относительно коэффициента корреляции. Нетрудно заметить, что центрирование и нормирование для линейной регрессии позволяет понизить на единицу размерность системы уравнений, т.е. упростить решение задачи определения коэффициентов, а самим коэффициентам придать ясный смысл.

Применение МНК для нелинейных функций практически ничем не отличается от рассмотренной схемы (только коэффициент a0 в исходном уравнении не равен нулю).

Например, пусть необходимо определить коэффициенты параболической регрессии



Выборочная дисперсия ошибки



На ее основе можно получить следующую систему уравнений



После преобразований система уравнений примет вид



Учитывая свойства моментов стандартизованных величин, запишем



Определение коэффициентов нелинейной регрессии основано на решении системы линейных уравнений. Для этого можно применять универсальные пакеты численных методов или специализированные пакеты обработки статистических данных.

С ростом степени уравнения регрессии возрастает и степень моментов распределения параметров, используемых для определения коэффициентов. Так, для определения коэффициентов уравнения регрессии второй степени используются моменты распределения параметров до четвертой степени включительно. Известно, что точность и достоверность оценки моментов по ограниченной выборке ЭД резко снижается с ростом их порядка. Применение в уравнениях регрессии полиномов степени выше второй нецелесообразно.

Качество полученного уравнения регрессии оценивают по степени близости между результатами наблюдений за показателем и предсказанными по уравнению регрессии значениями в заданных точках пространства параметров. Если результаты близки, то задачу регрессионного анализа можно считать решенной. В противном случае следует изменить уравнение регрессии (выбрать другую степень полинома или вообще другой тип уравнения) и повторить расчеты по оценке параметров.

При наличии нескольких показателей задача регрессионного анализа решается независимо для каждого из них.

Анализируя сущность уравнения регрессии, следует отметить следующие положения. Рассмотренный подход не обеспечивает раздельной (независимой) оценки коэффициентов – изменение значения одного коэффициента влечет изменение значений других. Полученные коэффициенты не следует рассматривать как вклад соответствующего параметра в значение показателя. Уравнение регрессии является всего лишь хорошим аналитическим описанием имеющихся ЭД, а не законом, описывающим взаимосвязи параметров и показателя. Это уравнение применяют для расчета значений показателя в заданном диапазоне изменения параметров. Оно ограниченно пригодно для расчета вне этого диапазона, т.е. его можно применять для решения задач интерполяции и в ограниченной степени для экстраполяции.

Главной причиной неточности прогноза является не столько неопределенность экстраполяции линии регрессии, сколько значительная вариация показателя за счет неучтенных в модели факторов. Ограничением возможности прогнозирования служит условие стабильности неучтенных в модели параметров и характера влияния учтенных факторов модели. Если резко меняется внешняя среда, то составленное уравнение регрессии потеряет свой смысл. Нельзя подставлять в уравнение регрессии такие значения факторов, которые значительно отличаются от представленных в ЭД. Рекомендуется не выходить за пределы одной трети размаха вариации параметра как за максимальное, так и за минимальное значения фактора.

Прогноз, полученный подстановкой в уравнение регрессии ожидаемого значения параметра, является точечным. Вероятность реализации такого прогноза ничтожна мала. Целесообразно определить доверительный интервал прогноза. Для индивидуальных значений показателя интервал должен учитывать ошибки в положении линии регрессии и отклонения индивидуальных значений от этой линии. Средняя ошибка прогноза показателя y для фактора х составит



где  – средняя ошибка положения линии регрессии в генеральной совокупности при *x* = *xk*;

 – оценка дисперсии отклонения показателя от линии регрессии в генеральной совокупности;

*xk* – ожидаемое значение фактора.

Доверительные границы прогноза, например, для уравнения регрессии (7.14), определяются выражением 

Отрицательная величина свободного члена *а0* в уравнении регрессии для исходных переменных означает, что область существования показателя не включает нулевых значений параметров. Если же  *а0 > 0*, то область существования показателя включает нулевые значения параметров, а сам коэффициент характеризует среднее значение показателя при отсутствии воздействий параметров.

*Задача 7.2.* Построить уравнение регрессии по выборке, заданной в табл. 10.2.

*Решение.* Применительно к указанной выборке построение аналитической зависимости в основной своей части выполнено в рамках корреляционного анализа: пропускная способность зависит только от параметра Ф1. Остается подставить в выражение (7.14) вычисленные ранее значения параметров. Уравнение для пропускной способности примет вид

*ŷ*= 26,47– 0,93×41,68×5,39/6,04+0,93×5,39/6,03×*х* = – 8,121+0,830*х*.

Результаты расчетов представлены в табл. 7.5.

Таблица 7.5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| N пп | Выходная характеристика | Ф1 | Значение функции | Погрешность |
| Y | X | ŷ | ε |
| 1 | 26.37 | 41.98 | 26.72 | -0.35 |
| 2 | 28.00 | 43.83 | 28.25 | -0.25 |
| 3 | 27/83 | 42.83 | 27.42 | 0.41 |
| 4 | 31.67 | 47.28 | 31.12 | 0.55 |
| 5 | 23.50 | 38.75 | 24.04 | -0.54 |
| 6 | 21.04 | 35.12 | 21.03 | 0.01 |
| 7 | 16.94 | 32.07 | 18.49 | -1.55 |
| 8 | 37.56 | 54.25 | 36.90 | 0.66 |
| 9 | 18.84 | 32.70 | 19.02 | -0.18 |
| 10 | 25.77 | 40.51 | 25.50 | 0.27 |
| 11 | 33.52 | 49.78 | 33.19 | 0.33 |
| 12 | 28.21 | 43.84 | 28.26 | -0.05 |
| 13 | 28.76 | 44.03 | 28.42 | 0.34 |
| 14 | 24.60 | 39.46 | 24.63 | -0.03 |
| 15 | 24.51 | 38.78 | 24.06 | 0.45 |

Остаточная дисперсия стандартизованной величины *Y* относительно37.56 стандартизованной величины *Х* равна 1– 0,932 = 0,14, т.е. является малой величиной. Погрешность аппроксимации и величина остаточной дисперсии показывают высокую точность линейной модели, поэтому задачу регрессионного анализа можно считать решенной. Свободный член уравнения регрессии отрицательный, следовательно, область существования показателя не включает нулевое значение параметра "отношение сигнал/шум", что вытекает из сущности параметра (при нулевом уровне сигнала передача информации невозможна).

Как уже было отмечено ранее, эффективное использование процедур статистического анализа экспериментальных данных возможно только в том случае, если в распоряжении исследователя имеются соответствующие инструментальные средства.

Основные итоги:

* При сравнении альтернативных решений одним из наиболее  
  эффективных инструментов их оценивания является математическое  
  моделирование.
* В свою очередь, в тех случаях, когда значения оценок зависят от  
  воздействия большого числа случайных факторов, либо интерес представляет  
  развитие ситуации во времени, удобнее всего использовать имитационные  
  модели.
* Создание имитационной модели сложной системы,  
  функционирование которой предполагает наличие параллельных  
  процессов, является весьма сложным делом, требующим от разработчика  
  не только хорошего знания рассматриваемой предметной области, но  
  достаточно прочных навыков в программировании.
* Результаты имитационного эксперимента могут быть использованы  
  для принятия решения лишь при условии их корректной статистической обработки, что предъявляет к уровню подготовки исследователя целый ряд дополнительных требований.

И, наконец, главный вывод. Существенное повышение технологичности подготовки, проведения и анализа результатов имитационного моделирования возможно в том случае, если в распоряжении исследователя имеются соответствующие инструментальные средства