МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ
УНИВЕРСИТЕТ

Алексей Сергеевич Дегтярев
Валерия Алексеевна Драбенко
Вадим Анатольевич Драбенко

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ МЕТЕОРОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

Учебник

Рекомендовано Учебно-методическим объединением ВУЗов России в качестве учебника для студентов высших учебных заведений

ООО «Андреевский издательский дом»
САНКТ-ПЕТЕРБУРГ, 2015

УДК (551.46+551+556): 519.22
ББК 26.23 А 79

Дегтярев А.С., Драбенко В.А., Драбенко В.А. Статистические методы обработки метеорологической информации. Учебник. - СПб: ООО «Андреевский издательский дом», 2015 - 225 с.

Рецензенты:
Васильев А.А., доктор физ.-мат.н., с.н.с., доцент кафедры физики атмосферы, Физический факультет, Санкт-Петербургский государственный университет.
Матвеев Ю.Л., доктор физ.-мат.н., профессор, заведующий кафедрой математического моделирования социально-экономических и природных процессов, Государственной полярной академии

Учебник написан для специалистов метеорологических специальностей и носит прикладной характер. Основной упор сделан на доступность изложения, для чего приводится большое количество примеров решения поставленных задач. Несмотря на некоторую потерю математической строгости изложения материала позволяет его легко усваивать студентам любого уровня подготовки.

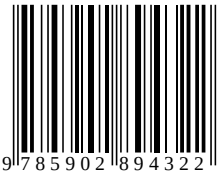
Рассматриваются различные функции риска, и вводится критерии оценки действия потребителя.

Все главы учебника посвящены рассмотрению вопросов построения прогностических методов и их оцениванию, как для потребителей общего назначения, так и для решения специализированных задач.

Учебник может быть полезен не только для студентов, но и магистрантов и аспирантов различных научных направлений, специализирующихся на применении прогностических методов.

Дегтярев А.С., Драбенко В.А., Драбенко В.А.
Статистические методы обработки метеорологической информации.
ООО «Андреевский издательский дом»
197738, Санкт-Петербург, пос. Репино, Приморское шоссе, д. 394
E-mail: **biom@nm.ru**
Подписано в печать 10.12.2014 г.
Печ. листов 11,2. Тираж 2000 экз.
Отпечатано с готовых диапозитивов в РГГМУ

ISBN 978-5-902894-32-2



ПРЕДИСЛОВИЕ

Обработка метеорологической информации является необходимым этапом принятия решения большинством потребителей этой информации. Обработка информации осуществляется непосредственно специалистами метеослужбы, или самими потребителями.

В настоящее время в схеме обработки метеорологической информации можно выделить два принципиально разных подхода: гидродинамический и физико-статистический.

Настоящий учебник посвящен только одному подходу к проблемам обработки метеорологической информации – физико-статистическому или просто статистическому, и предназначен, прежде всего, для подготовки метеорологов, гидрологов и океанологов. Так как основной задачей метеоролога является разработка прогноза погоды, то и в учебнике рассматриваются те методы математической статистики, которые позволяют строить прогностические алгоритмы и оценивать качество.

Специфика подготовки инженеров-метеорологов состоит в том, что по уровню математического образования они должны приближаться к математикам, а по уровню обслуживания потребителя должны быть опытными практиками.

В настоящем учебнике авторы излагают современный аппарат математической статистики на языке, доступном обычному инженеру-синоптику. Упор сделан на изучение принципов решения задачи, на рассмотрении основных подходов к построению алгоритмов. Авторами практически не приводится теорем и их доказательств, так как при необходимости они легко могут быть найдены в специальной литературе.

Отличительной особенностью данного учебника является то, что многие задачи статистической обработки информации рассматриваются с позиций оптимального использования информации потребителем, предлагаются новые критерии оценки эффективности метеорологической информации.

1 ПРОБЛЕМА СТАТИСТИЧЕСКОЙ ОБРАБОТКИ МЕТЕОРОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

1.1 Классификация видов метеорологической информации

В настоящее время трудно назвать сферу человеческой деятельности, которая не зависела бы от погодных условий. Сельское хозяйство и транспорт, рыболовство и строительство, энергетика и лесное хозяйство – вот только часть отраслей народного хозяйства, подверженных влиянию атмосферных факторов. По оценкам различных авторов годовой экономический эффект от деятельности метеорологической службы составляет по нашей стране около одного миллиарда долларов.

Особое значение метеорологические условия имеют для авиации. Высота облаков и дальность видимости в пункте взлета и посадки, конвективные явления, турбулентность, скорость и направление ветра на маршруте полета – вот далеко не полный перечень метеорологических величин, влияющих на безопасность и экономическую эффективность полетов.

В процессе своей деятельности потребители пользуются тремя основными видами метеорологической информации: первичной, прогностической и климатологической (рис. 1.1)

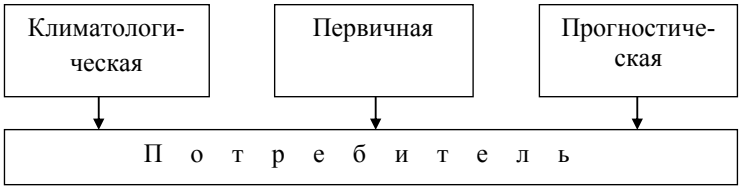


Рис. 1.1 Виды метеорологической информации, передаваемой потребителю

Первичная метеорологическая информации, т. е. информация об исходном состоянии атмосферы, включает в себя результаты наблюдений за физическими процессами, происходящими в атмосфере и на подстилающей поверхности. Наблюдение является одним из основных способов изучения атмосферы. На основе наблюдений метеоролог строит физические модели атмосферных процессов, разрабатывает прогнозы погоды, дает рекомендации потребителю метеорологической информации на принятие решения. Наблюдение сопровождается приемом, преобразованием и регистрацией информации о свойствах наблюдаемого процесса.

Наблюдение за погодой проводится как с помощью органов чувств, так и

с помощью приборов (термометров, барометров, радиолокаторов и др.). регистрация информации происходит как в мозгу человека, так и на специальных носителях (бумаге, магнитных дисках, фотопленке и т. д.).

До принятия решения информация, получаемая в результате наблюдения, претерпевает, как правило, существенные преобразования. Наиболее характерными этапами таких преобразований в большинстве сфер человеческой деятельности являются измерения и обработка результатов измерений.

Измерение состоит в сравнении наблюдаемой величины (свойства объекта) с эталоном и получении в результате этого ее численного значения. При визуальном измерении дальности видимости наблюдатель определяет самый удаленный видимый объект и, пользуясь схемой видимости ориентиров, определяет видимость. При приборных измерениях регистрируется значения измеряемого параметра.

Значений большинства наблюдений является регистрация определенных параметров, характеризующих изучаемые явления или процесс. Эти параметры будем обозначать через $y_1, y_2, \ldots, y_n$. Очевидно, что возможны два случая измерений.

В первом случае измеряются непосредственно интересующие нас величины $y_1, y_2, \ldots, y_n$. Например, измеряется линейкой расстояние между двумя точками. Тогда говорят, что имеют место прямые измерения.

Во втором случае, величины $y_1, y_2, \ldots, y_n$ непосредственно измерены быть не могут, а измеряются другие физические величины $x_1, x_2, \ldots, x_n$, с которыми связаны величины $y_1, y_2, \ldots, y_n$. Такие измерения называют косвенными. Например, плотность воздуха ρ сама по себе не измеряется, но она может быть рассчитана через значения измеренного давления P и температура воздуха T по формуле:

$$\rho = \frac{P}{RT},$$

где R – универсальная газовая постоянная.

Результат измерения всегда содержит ошибку (погрешность), представляющую собой отклонение результата измерения от истинного значения измеряемой величины.

Данное утверждение не всегда относится к наблюдениям, основанным на непосредственном подсчете. Так, например, синоптик может точно определить количество циклонов, проходящих через рассматриваемый район или количество гроз в районе аэродрома в течение некоторого периода времени. В тоже время в определенных условиях (дефицит времени, усталость, плохие условия наблюдения) наблюдатель может допустить ошибки при подсчете.

Существующая в настоящее время теория ошибок измерений используется

в основном для оценки непрерывных физических величии. При этом ошибки измерений делятся на систематические и случайные.

Систематические ошибки вызываются причинами, действующими одинаково при многократных измерениях. Так, например, если капилляр термометра имеет больший диаметр, чем положено, то он будет показывать все время значения температуры ниже истинных. Систематические ошибки при повторных измерениях остаются неизменными, а если изменяются, то закономерно. Систематические ошибки, как правило, обусловлены погрешностями измерительных приборов (инструментальные ошибки) и несовершенством методов измерений (методические ошибки).

По своей природе и характеру результаты наблюдений, представленные в любой форме (числовой, графической и т. д.), являются чаще всего случайными, что связано как с вероятной природой изучаемого явления, так и с различными случайными возмущениями, которые неизбежно сопровождают сам процесс наблюдения. Поэтому даже в простейших экспериментах причинно-следственная связь между отдельными сторонами и компонентами данного явления оказывается настолько «затушенной» и неочевидной, что установить ее удается лишь после тщательного и кропотливого анализа с применением достаточного совершенного математического аппарата.

В настоящем учебнике рассматриваются методы обработки наблюдений, имеющих случайные ошибки.

Прогностическая информация содержит в себе сведения об ожидаемых значениях метеорологических величин и явлениях. Она получается в результате обработки первичной метеорологической информации на основе применения некоторых моделей (рис. 1.2).

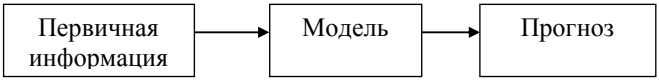


Рис. 1.2. Схема получения прогностической информации

Целью применения модели является установление связи между прогнозируемыми величинами y и первичной метеорологической информацией. Естественно, что устанавливаемая связь будет отличаться от реальной. Это отличие, также как и в предыдущем случае, возникает за счет систематических и случайных ошибок, что, в свою очередь, приводит к наличию ошибок в природе.

Климатологическая информации содержит в себе сведения о наиболее часто повторяющихся значениях метеорологических элементов, о их максимальных и минимальных значениях, о законах распределения элементов и явлений во

времени и пространстве и т. д. Так как эти характеристики получаются на основе обработки первичной информации, то в них также содержатся ошибки. Кроме того, появляются ошибки, связанные с объемом архивной выборки.

1.2 Математические модели обработки информации

Математической моделью обработки информации называется совокупность математических зависимостей, привлекаемых для описания изучаемого процесса или явления. С помощью модели можно установить связи между измеряемыми величинами, между наблюдаемыми и прогнозируемыми характеристиками состояния атмосферы и т. д.

Математические модели основываются на использовании двух видов связей между величинами X и Y: функциональной и статистической. Функциональная связь устанавливает однозначное соответствие: каждому значению x соответствует одно и только одно значение y. Статистическая связь устанавливает неоднозначное соответствие между величинами: каждому значению x ставится в соответствие одно и только одно распределение повторяемостей значения y. Эти определения будут полностью соответствовать и векторам $X_{<n>}$ и $Y_{<m>}$.

В связи с тем, что в метеорологии функциональные связи чаще всего описываются с помощью уравнений гидротермодинамики, модели, основанные на их использовании, называют гидродинамическими.

Гидродинамические модели широко используются при разработке прогнозов полей метеорологических элементов: давления, геопотенциала, температуры, ветра и др. Их достоинством является получение точного результата в случае соответствия состояния системы начальным и граничным условиям модели. Если же такого соответствия не наблюдается, то использование гидродинамической модели может привести к прямо противоположному результату. Кроме того, при большом количестве уравнений, входящих в систему, в результате округлений происходит накопление ошибок вычислений, что приводит к ошибочным результатам. Чем больше уравнений, тем больше ошибки. Точные решения систем уравнений на основе использования метода Гаусса возможны, только если число уравнений не превышает 100. Если число этих уравнений находится в пределах от 100 до 1000, то используются приближенные методы. В случае, когда это число превышает 1000, возможны решения только с помощью статистических моделей.

Таким образом, статистические методы еще длительное время будут являться основным инструментом деятельности Гидрометслужбы при обеспечении различных потребителей.

Стоит также отметить, что использование для замыкания системы уравнения состояния, также вносит дополнительные погрешности вычислений, потому

что это уравнение справедливо только для замкнутых систем, а атмосфера таковой не является.

Появление мощных вычислительных комплексов в Москве, Новосибирске и Хабаровске позволило существенно повысить качество гидродинамических прогнозов. Однако, как отмечалось на VII Всероссийском метеорологическом съезде, качество статистических моделей и, следовательно, прогнозов пока превышает качество гидродинамических.

Отметим, что в силу стохастической (вероятностной) природы первичной метеорологической информации, результаты, получаемые с помощью гидродинамических моделей, также будут носить стохастический характер.

Математические модели, основанные на использовании статистических связей, называются статистическими или физико-статистическими моделями. Из анализа видов метеорологической информации легко сделать вывод о том, что статистические модели имеют очень большое значение при обеспечении потребителей метеорологической информацией. Они играют исключительно важную роль в процессе познания окружающей нас действительности, помогают вскрывать объективные закономерности развития и взаимосвязи предметов и явлений, широко используются при анализе состояния дел в экономике и науке, при разработке планов и долгосрочных программ развития народного хозяйства и Вооруженных Сил.

Применительно к прогнозу статистическую модель связи между векторами $Y_{<m>}$ и $X_{<n>}$ можно представить как соотношение:

$$Y_{<m>} = \mathrm{F}(X)_{<n>} + \varepsilon, \qquad (1.1)$$

где $X_{<n>}$ – вектор n исходных характеристик состояния атмосферы,

$Y_{<m>}$ – вектор прогнозируемых характеристик,

ε – погрешность в определении вектора $Y_{<m>}$.

Применительно к теории измерений в (1.1) $Y_{<m>}$ - наблюдаемый сигнал, $\mathrm{F}(X)_{<n>}$ – полезная составляющая сигнала наблюдений, ε – помеха.

В статистических моделях, так же как и в гидродинамических, основной задачей является установление формы связи $\mathrm{F}(X_{<n>})$, но дополнительно возникает еще одна задача – определение ε, причем кроме величины ε требуется устанавливать и закон ее распределения и выяснить причины возникновения.

В данном курсе будут рассматриваться только статистические модели.

1.3. Общая схема обработки метеорологической информации

Смысл обработки метеорологической информации заключается в выдаче рекомендаций потребителю на принятие решения. Потребителем может являться

хозяйственник, командир, летный состав, научные работники и сами метеорологи. Каждому из этих потребителей необходим свой перечень сведений о состоянии атмосферы и своя форма представления информации.

Хозяйственнику, командиру необходимо знать – сможет ли он по метеорологическим условиям выполнить поставленную задачу в установленные сроки и, если сможет, то какими силами это лучше сделать.

Синоптику для разработки прогноза необходимо иметь данные приземных, аэрологических, спутниковых, радиолокационных измерений и наблюдений. Специалистам по активным воздействиям, кроме того, требуются сведения о химическом составе атмосферы.

В зависимости от решаемых задач потребителю представляется один из трех видов метеорологической информации: первичная, прогностическая или климатологическая.

Первичная информация выдается непосредственно перед моментом принятия решения. При заходе на посадку летчику необходимо знать давление на взлетно-посадочной полосе для определения высоты полета, скорость боковой составляющей ветра, высоту нижней границы облаков, дальность видимости.

Если облачность не позволяет совершать посадку, то можно попытаться на нее воздействовать. Решение на активное воздействие принимается с учетом необходимости выполнения задачи, наличия соответствующих реагентов, экономических затрат и экологических последствий.

В первичной информации нуждаются синоптики, климатологи, специалисты по численным прогнозам.

Прогностическая информация, т. е. информация о будущем состоянии атмосферы, необходима для принятия решения в некоторый период времени (классификация прогнозов дается в специальных курсах). В ней нуждаются практически все потребители. Но для разработки прогноза состояния атмосферы необходимо строить сначала физическую модель процесса, затем модель принятия решения на формулировку прогноза погоды и, только в последнюю очередь, дается прогноз выполнения задачи по метеоусловиям.

Климатологическая информация, также как и прогностическая, прежде чем попасть к потребителю должна проходить несколько этапов обработки. На первом этапе в результате обработки первичной информации формируется собственно банк климатологической информации. Затем, в соответствии с целевыми функциями потребителя из этого банка строится банк специализированной климатологической информации. На следующем этапе рассчитываются вероятности выполнении задачи и выдаются рекомендации на ее выполнение.

Общая схема обработки метеорологической информации приведена на

рис. 1.3.

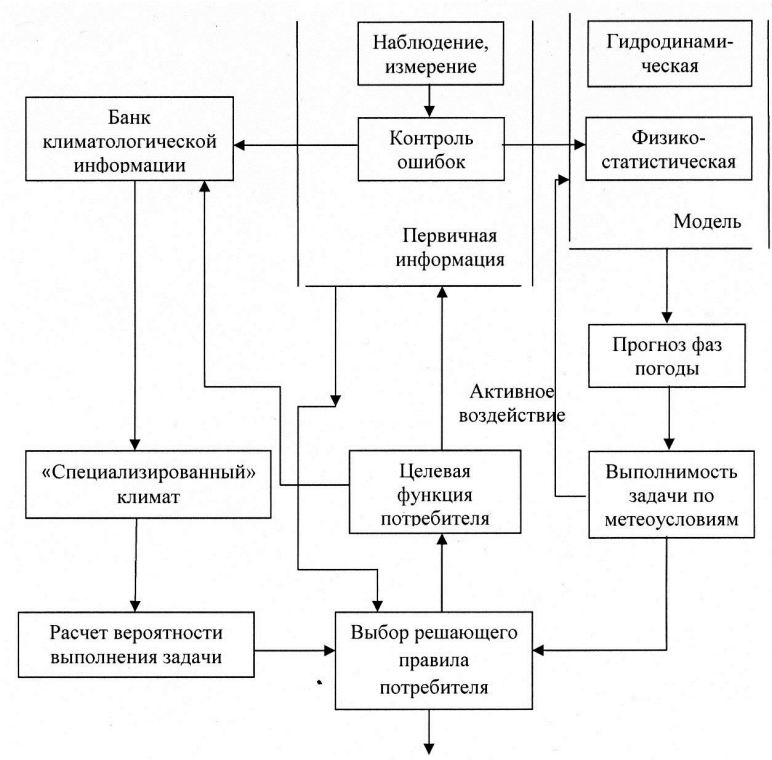


Рис. 1.3. Общая схема обработки метеорологической информации

1.4 Задачи и методы статистической обработки метеорологической информации

Основной целью обработки метеорологической информации является получение сведений об интересующем нас объекте. Особенности процесса обработки зависят от характера решаемых задач, объема используемой информации, оперативности обработки. По этим признакам обработка может подразделяться на первичную и вторичную, полную и неполную, оперативную и неоперативную.

Задачами первичной обработки является выделение полезного сигнала на фоне помех, сжатие информации и приведение ее к наиболее удобному для дальнейшей обработки виду.

Один из возможных вариантов перечня задач, решаемых при первичной и вторичной обработке, представлен на рис 1.4.

Рис. 1.4. Возможный вариант перечня задач, решаемых при первичной и вторичной обработке информации.

Наиболее характерной операцией выделения полезного сигнала является устранение грубых ошибок. Эта операция получила название контроля. Из рисунка видно, что на этапе первичной обработки результаты наблюдений «очищаются», «сжимаются» и канонизируются. После этого они могут быть подвержены дальнейшей обработке или занесены в соответствующие архивы. Наиболее часто статистические методы используются для решения задач выделения полезного сигнала и устранения грубых ошибок.

Задачи второй обработки информации могут быть сведены и две больших группы – построения математических моделей реальных процессов и анализа этих моделей.

Наиболее широко статистические методы применяются при построении математических моделей, как закон распределения, уравнения регрессии, дискриминантные функции и т. д. типовыми законами построения этих моделей являются задачи оценивания параметров распределения и законов распределения, числовых характеристик случайных величин и функций, проверка различных гипотез. В метеорологии, особенно в синоптике, особая роль отводится построению различных прогностических связей на основе регрессионного анализа и теории распознания образов.

В задачах анализа моделей и исходной информации производится оценка влияния различных факторов на рассматриваемый процесс и выбор наиболее важных из них, исследуется внутренняя структура результатов наблюдений и построенных моделей и т. д.

По условиям использования статистическая обработка подразделяется на оперативную и неоперативную.

Оперативная обработка производится непосредственно по мере поступления информации и осуществляется с целью принятии наиболее важных на данный период решений. К таким решениям относятся: выдача информации о погоде в пункте посадки, разработка прогнозов погоды на полеты по существующим алгоритмам и т. д.

Неоперативная обработка предназначена для использования информации в течение длительных промежутков времени. К такому виду обработки относятся: разработка авиационно-климатических описаний и справок, построение различных прогностических алгоритмов и оценка их эффективности и т. д.

Естественно, что оперативная и неоперативная обработка требуют различных объемов информации. Для разработки прогностического алгоритма необходимо всесторонне проанализировать изучаемый процесс, выбрать наиболее влияющие на него факторы, построить прогностические связи и оценить их и т. д. Поэтому здесь используется вся информация, получаемая в процессе наблюдения, а сама обработка называется полной. При разработке прогнозов по имеющимся алгоритмам требуется выбрать наиболее подходящий для данной ситуации способ и уже для этого способа взять необходимую информацию. В этом случае используется только часть информации о состоянии объектов, и такая обработка называется неполной. В связи с вышеизложенным, часто отождествляют понятия оперативной и неполной, неоперативной и полной обработкой.

Все эти задачи решаются с привлечением различных методов математической статистики. Сам термин «статистика» происходит от латинского слова «status», означающего «положение», «состояние». Задолго до возникновения статистики как науки в древнейших и средневековых государствах выполнялась обработка результатов наблюдений, получаемых при переписи населения, определения его имущественного положения, размеров налогообложения, объема производимой сельскохозяйственной продукции и т. д.

Как наука статистика зародилась в Англии во второй половине XYII в. В работах видных экономистов той эпохи Д. Гроунта и У. Петти подчеркивалось, что статистика – это не только собирание и подытоживание фактов и сведений, но и совокупность методов, позволяющих выявить конкретные социальные и

экономические закономерности, определенные тенденции развития общественных явлений и процессов.

В конце XYIII в. в связи с необходимостью учета случайного характера процесса измерений получила развитие теория ошибок. Разработка статистических методов длительное время стимулировалась потребностями этой теории. В начале XIX в. новый импульс в развитии статистики и выделении ее в самостоятельную математическую дисциплину дали биологические исследования, в частности потребности медицины и генетики. В это время начала формироваться отдельная дисциплина, изучающая методы статистической обработки, которая получила название «математическая статистика».

В середине XIX в. Начала формироваться и наука, занимающаяся изучением неопределенностей в природе, которая получила название «теория вероятностей». Эти две дисциплины неразрывно связаны между собой: у них общая терминология, общий математический аппарат. Основное их отличие состоит в том, что математическая статистика имеет дело с ограниченными выборками, а теория вероятностей – с неограниченными. Это различие приводит к разным методам решения задачи.

Основное теоретическое содержание классической математической статистики составляет три ее раздела – методы оценивания параметров законов распределения и самих законов распределения, методы проверки статистических гипотез и методы построения и анализа статистических зависимостей. В настоящее время многие методы в математической статистике обобщаются с позицией теории статистических решений, которая позволяет в наиболее общем виде формировать и решать большинство статистических задач, в том числе и задач оценивания параметров и проверки гипотез. В то же время методы проверки гипотез могут быть приложены к получению оценок параметров распределения.

Для построения и анализа зависимостей широко применяются такие методы статистической обработки, как дисперсионный и корреляционной, компонентный и факторный, регрессионный, кластерный и дискриминантный анализы и т. д. Перечисление теории и методы базируются в свою очередь на выборочном методе. Целесообразность и эффективность применения того или иного метода математической статистики для обработки метеорологической информации определяются многими факторами. Основные из них – характер задачи обработки, природа наблюдаемого объекта, объем выборки или измерений (прямые или косвенные), объем априорной информации. Эти факторы и могут быть положены в основу классификации статистических методов обработки информации.

В зависимости от объема априорной информации методы делятся на параметрические (известны законы распределения с точностью до параметров) и непараметрические (законы распределений неизвестны). По объему выборки различают методы обработки при «большой» выборке и методы обработки при «малой» выборке. В зависимости от типа измерений используются методы обработки результатов прямых измерений и косвенных. Свойства обрабатываемых данных привели к развитию методов обработки скалярных и векторных случайных объектов, случайных величин и функций, стационарных и нестационарных случайных функций и т. д.

С 60-х годов началось интенсивное внедрение методов математической статистики в метеорологию. В результате проведенных исследований удалось создать теорию построения оценки прогностических алгоритмов, внедрить в практику метеорологического обеспечения вероятностные прогнозы, получить возможность качественного исследования стационарных случайных процессов и т. д.

2 ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЫБОРОК

2.1. Сущность выборочного метода, генеральная и выборочная совокупность

Математическая статистика как наука занимается исследованиями природных процессов на основе анализа экспериментальных данных. Причем исследования проводятся в массовых случайных явлениях и предполагается, что опыты могут быть повторены большое число раз при одних и тех же условиях. В результате проведения каждого опыта регистрируется определенный признак изучаемого объекта. Различают общий и основной признаки. **Общим признаком** называется свойство, по которому объекты объединяются в однородные совокупности, а **основным признаком** – свойство объектов, исследуемое в данном эксперименте.

Пример 1. Требуется исследовать возможность образования метелей в «ныряющих» циклонах. Для этой задачи общим признаком является траектория движения циклона, по которой определяется сам факт «ныряния». Выбираются все циклоны, имеющие соответствующие траектории. Затем в каждом из таких циклонов рассматриваются значения скоростей ветра, интенсивность снега и т. д. Выбранные признаки будут являться основными.

Под однородной совокупностью будем понимать совокупность, все элементы которой являются реализациями одной и той же случайной величины (функции) с одним и тем же законом распределения. В нашем примере такой реализацией будет траектория движения циклона, направленная с севера-запада на юго-восток или с севера на юг.

Общие и основные признаки могут быть как количественными (скорость и направление ветра, температура, количество облачности в баллах и т. д.) так и качественными (форма облаков, тип циклона, наличие или отсутствие опасного явления погоды и т. д.). в математической статистике отдельное конкретное значение наблюдаемого основного признака называют его реализацией или вариантом.

Совокупность наблюдаемых реализаций называется выборкой. Из закона больших чисел следует, что с увеличением числа испытаний точность и надежность получаемых результатов увеличивается, и в пределе результаты стремятся к своим истинным значениям. Отсюда следует вывод, что чем больше выборка, тем лучше результат. Бесконечную (гипотетическую) совокупность возможных реализаций принято называть генеральной совокупностью. Это понятие представляет собой идеализацию действительного положения вещей, даже когда под

этим понимается потенциальная возможность неограниченного повторения опытов. На практике под генеральной совокупностью (выборкой) понимают всякую выборку, «достаточно большую» по сравнению с объемом имеющейся выборки.

При статистических исследованиях вероятностных свойств совокупности объектов практически невозможно произвести опыты над каждым объектом. Для определения состояния зерновых перед посевом невозможно обследовать каждое зерно, при оценке боевых качеств ракет невозможно провести их общий отстрел, при проверке работы радиозондов невозможно их все запустить и т. д. Вместе с тем необходимо получить эти данные, для чего и разработан выборочный метод.

Суть выборочного метода состоит в том, что основные свойства и признаки генеральной совокупности изучаются по некоторой ее части, называемой выборочной совокупности.

Пример 2. Рассмотрим набор (группу или партию) более или менее схожих предметов, состоящих из отдельных единиц, которые отличаются друг от друга по определенному признаку. Пусть это будут обработанные бруски длиной 500 мм. За счет флуктуаций в процессе производства длина каждого бруска в отдельности будет отличаться от этого значения, но не должна выходить из диапазона 495 – 505 мм. Бруски, длина которых выходит из этого диапазона, считаются бракованными. Требуется определить долю годных брусков во всей выпущенной партии 10000 штук. Естественно, что по практическим соображениям все бруски измерять нецелесообразно. Вместо этого можно проверить выборку из брусков, определив заранее ее объем, например 100 штук.

Выборочную совокупность часто называют статистической совокупностью или рядом наблюдений. Выборку, элементами которой являются метеорологические наблюдения, принято называть метеорологическим рядом наблюдений. Число наблюдений в выборке называется ее объемом и, как правило, обозначается через n. В нашем примере объем выборки равен 100.

Заметим, что выборка является однородной, если все ее элементы извлечены из одной генеральной совокупности.

В математической статистике предполагается, что выборки формируются при многократной реализации случайного эксперимента, результат которого нельзя заранее точно предсказать. Поэтому такие выборки называют случайными.

Пусть количественно исследуемый основной признак описывается случайной величиной $\hat{X}$. Допустим, что в процессе наблюдений основного признака получена последовательность n значений $x_1, x_2, \ldots, x_n$ случайной величины $\hat{X}$. До проведения эксперимента эта последовательность является случайной выборкой

и обозначается$< \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n >$.

Если основной признак наблюдаемого объекта описывается случайной функцией $\hat{X}$ (t), то в процессе наблюдений получаем конечную совокупность (ансамбль) реализаций $x(t)_1, x(t)_2, \ldots, x(t)_n$, t[0,m] случайной функции. Модель случайно выборки представляет в этом случае последовательность случайных функций

$$< \hat{x}(t)_1, \hat{x}(t)_2, \ldots, \hat{x}(t)_n >.$$

Иногда выборку удобно описывать с помощью n-мерного случайного вектора $\hat{X}_{<n>}$, компонентами которого являются элементы последовательности $< \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n >$.

Случайный характер данной выборки выражается в том, что нельзя заранее предсказать возможные значения элементов выборки, и любые две последовательности n наблюдаемых значений величины X в общем случае будут различными. В связи с этим наши выводы о вероятностных характеристиках исследуемой генеральной совокупности будут изменяться при переходе от одной случайной выборки к другой. Возникает вопрос о репрезентативности (представительности) выборки. Выборка должна быть организована таким образом, чтобы она наилучшим образом представляла структуру (распределение) исследуемой генеральной совокупности. Она должна содержать максимально возможную информацию об этой совокупности.

Для выполнения этого требования необходимо каждый элемент выборки отбирать независимо друг от друга и случайным образом, т. е. так, чтобы любой элемент генеральной совокупности имел равные шансы попасть в формируемую выборку. В конкретных прикладных задачах элементы выборок представляют собой реализации случайных величин, т. е. детерминированные величины. Таким образом, априорно (до проведении опыта) выборка будет случайной, а апостериорно (после проведения опыта) – неслучайной.

В статистике различают повторные и бесповторные выборки. Выборка называется повторной, если отобранный элемент после испытания снова возвращается в генеральную совокупность до следующего испытания. Выборка называется бесповторной, если отобранный объект после испытания не возвращается в генеральную совокупность.

Если выбираемые элементы извлекаются по одному из всей генеральной совокупности, то полученная выборка называется простой*.

Если исследователя интересует несколько основных признаков, например высота нижней и верхней границ облаков, то экспериментатор наблюдает векторную случайную величину, которую в общем случае будем обозначать через

$\hat{X}_{<q>}$. При этом выборка $\langle < \hat{x}_{<q>1}, \hat{x}_{<q>2}, \dots, \hat{x}_{<q>n} > \rangle$ описывает nмерным случайным вектором $\hat{X}_{<nq>}$.

В задачах статистической обработки метеорологической информации принцип случайного отбора элементов выборки осуществить практически невозможно.

Метеорологические ряды формируются, как правило, на основе результатов последовательных или почти последовательных наблюдений за метеорологическими явлениями и величинами. Это приводит к тому, что элементы выборки становятся зависимыми друг от друга и не являются случайными. Так, например, при построении способа прогноза гроз разработчик способа в выборку включает все случаи, когда наблюдалась гроза, и случаи возникновения грозы за несколько дней подряд. Считать их независимыми для внутримассовой обстановки нельзя, так как каждая предыдущая гроза соответствующим образом воздействует на атмосферу и облегчает развитие последующей.

Считается, что в метеорологических рядах лишь 1/3 наблюдений оказывается практически взаимно независимыми. Причем, чем меньше временной интервал между наблюдениями, тем в большей степени элементы выборки зависят друг от друга. За счет этого теряется информативность выборки. При таком объективно существующем недостатке метеорологических рядов для получения по ним надежных выводов о статистической структуре исследуемых метеорологических величин необходимо, чтобы их объем был по возможности больше.

При использовании выборочного метода приходится всегда решать одну сложную, но практически очень важную задачу – определение минимального объема выборки, позволяющего обеспечить достаточно надежные выводы по исследуемому вопросу. В каждом конкретном случае, в зависимости от содержания рассматриваемой задачи и возможностей получения исходной информации, эта проблема может решаться по-разному. Однако, в любом случае при определении минимально необходимого объема выборки нужно как можно полнее учитывать следующие определяющие его факторы: требуемую точность и надежность анализируемых характеристик и их особенности; степень рассеивания и предполагаемый вид кривой распределения исследуемой случайно величины и в какой мере удается обеспечить независимость и случайность отбора элементов используемой выборки.

Минимально необходимый объем выборки является функцией тех характеристик случайной величины, точность и надежность определения которых, в свою очередь, зависит от объема выборки. В этом состоит основная трудность его определения, которая усугубляется тем. Что при обработке метеорологиче-

ской информации определить степень взаимной независимости элементов формируемой выборки очень сложно.

Выборки по объему могут быть как «большими», так и «малыми». Понятия «большой « и «малый» объем в математической статистике во многом условны. При практических расчетах некоторые авторы считают, что выборка является «малой», если число реализаций менее 50. Однако для метеорологических задач, в которых рассматриваются многомерные составляющие, это число никоим образом, не соответствует предъявляемым требованиям. Более логично в определении «большого» и «малого» объема выборки является следующий подход. При обработке метеорологических рядов производится группирование элементов выборки (см. подраздел 2.2). при этом количество информации, извлекаемой из выборки обычно уменьшается, что, приводит к ухудшению качества (точности, надежности и т. д.) получаемых результатов.

Выборку считают малой, если при ее обработке методами, основанными на группировании элементов, нельзя достичь заданного качества результатов. Выборка является большой, если она допускает группирование элементов без заметной потери количества информации.

Случайные выборки используются для решения указанных ранее задач математической статистики. При этом, как правило, элементы выборки используются для образования по соответствующим правилам новых случайных величин вида $\hat{S}_i = f_i\,(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$, которые называются статистиками.

Примерами статистик являются: выборочная сумма $\hat{S}_1 = \sum_{i=1}^{n} \mathrm{x_i}$, выборочное среднее $\hat{S}_2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x_i}$ и т. д.

Все свойства (характеристики) генеральной совокупности, получаемые по данным выборки, называются выборочными или статистическими. Эти же свойства, изучаемые в теории вероятностей, называют теоретическими. Отличие теоретических характеристик от статистических состоит в том, что теория не доказывает и не может доказать независимость элементов выборки.

2.2. Статистический ряд распределения и порядок его построения

Первым шагом на пути статистической обработки результатов наблюдений случайных объектов сбор статистических данных.

Простейшей формой записи результатов метеорологических наблюдений служит простой статистический ряд, который обычно оформляется в виде таблицы последовательных записей. Форма такой таблицы в каждом конкретном случае выбирается исходя из удобства первичной записи и последующей обработки результатов наблюдений. В качестве примера простого статистического ряда может служить дневник погоды (табл. 2.1).

Таблица 2.1

Фрагмент дневника погоды по станции Воронеж за 30 мая 1991 г.

Сроки наблюдения	Облачность			Ветер		Горизонтальная видимость, км	Температура воздуха, °С	Температура смоченного термометра, °С	Точка росы, °С	Влажность, %	Давление на уровне ВПП, м/б	Величина и характеристика барической тенденции
	Количество баллов	Форма	Высота нижней границы, м	Направление, град	Скорость, м/с							
1	2	3	4	5	6	7	8	9	10	11	12	13
01	9/9	Ns	1000	270	1	10	15.8	15.2	14.8	94	993.4	-0.5
02	9/9	Ns	1000	300	2	10	15.4	15.0	14.7	95	993.4	0.0
03	10/10	Sc	2200	320	3	6	15.4	14.4	13.6	89	990.7	-2.7
04	10/10	Sc	600	320	2	4	14.3	13.5	12.8	91	990.6	-0.1
05	10/10	Sc	600	320	3	4	13.9	13.1	12.4	96	993.1	+2.5
06	3/2	Cu	1000	320	4	6	13.2	12.9	12.6	96	992.5	-0.6
07	10/0	Ci	-	340	8	6	12.5	11.9	11.3	92	992.4	-0.1

При изучении распределений метеорологических случайных величин, которые, как правило, являются непрерывными, приходится прибегать к большому числу наблюдений (обычно несколько сотен случаев). В таких ситуациях выявить характер распределений случайной величины непосредственно из простого статистического ряда наблюдений практически невозможно. Поэтому он подвергается соответствующей обработке. Первым шагом такой обработки является формирование рядов, содержащих только необходимую исследователю информацию. Так, например, если требуется изучить распределение температуры, то из всей Табл. 2.1 выбираются только 1 и 8-й столбцы и строится таблица 2.2, которая представляет собой простой статистический ряд, но более удобный для изучения температурного режима.

Таблица 2.2

Фрагменты распределения температуры воздуха по станции Воронеж за 30 мая 1991 г.

Срок наблюдения	01	02	03	04	05	06	07	…
Температура воздуха	15,8	15,4	15,4	14,3	13,9	13,2	12,5	…

Видно, что если в этом ряду будет несколько сотен случаев, то сделать какие-либо выводы о температурном режиме достаточно сложно. Для упрощения

анализа простой статистический ряд преобразуется в вариационный или ранжированный ряд.

Вариационным (ранжированным) статистическим рядом называется ряд, в котором элементы выборки расположены в порядке возрастания: 12,5; 13,2; 13,9; 14,3; 15,4; 15,4; 15,8 °С.

Вариационный ряд, упорядочен по сравнению с простым статистическим рядом, но неудобен для анализа из-за большого количества случаев наблюдения.

$$H_1D_1,\ H_1D_2,\quad H_1D_3\ \ldots,\ H_1D_{10}$$

$$\ldots,\ \ldots,\ \ldots,\ \ldots$$

$$H_{10}D_1,\ H_{10}D_2,\ H_{10}D_3,\ \ldots, H_{10}D_{10}$$

Таким образом, получается 100 возможных сочетаний градаций случайных величин ($\widehat{H}$) и ($\widehat{D}$). Так как в архивной выборке всего 100 случаев, то в большую часть сочетаний градаций не попадает ни одного случая (при равномерном законе по одному), а в некоторые – по одному или по два. Естественно, что по одному случаю делать выводы о вероятностных характеристиках и соответствующих законах распределения нецелесообразно. Считается, что о вероятности можно судить тогда, когда имеется хотя бы 10 случаев попадания в градацию. Исходя из этих соображений, необходимый объем выборки должен определиться по формуле n = 10 t, где t – число возможных сочетаний градаций случайных величин. Так как t равно произведению градаций всех случайных величин, то

$$\text{n} = 10 \qquad k_1 \quad k_2 \quad \ldots . k_n \tag{2.3}$$

где k_1 – количество градаций 1-ой случайной величины;

k_2 – количество градаций2-ой случайной величины и т. д.

–

Таким образом, зная имеющийся объем архивной выборки, с помощью формулы 2.3 можно подобрать количество градаций, на которое разбивается каждая случайная величина.

По выбранному числу градаций k и значению размаха выборки L находим минимальную ширину градаций ΔX.

$$\Delta \text{X} = \frac{L}{k} = \frac{X_{max} - X_{min}}{k}. \tag{2.4}$$

Полученное по этой формуле значение ΔX может оказаться не совсем удобным для построения ряда числом. Так, например, ширина градации температуры может оказаться равной 2,43 °. В этом случае проводим округление в большую сторону т. е. до 3. В меньшую сторону не округляем, так как, при этом число

градаций окажется больше допустимого. После этой операции выбирается положение левой границы первой градации ряда распределения x_1 из условия $x_1 \leq x_{min}$ и с таким расчетом, чтобы положение правой границы его последней градации x_{k+1} соответствовало условия

$x_{k+1} \geq x_{max}$. По возможности значения границ градаций следует быть равными целому числу, что упрощает оформление и наглядность при практическом использовании полученного статистического ряда распределения. В процессе группирования выборки могут быть случаи точного совпадения значений случайной величины с границами градаций. Возникает вопрос: куда отнести данный случай? В такой ситуации можно поступить двумя способами. Во-первых, границы градации можно задавать числами с большим количеством значащих цифр, чем точность измерения членов обрабатываемого ряда. Например, если температура измерена с точностью до 0,1 °С, то границу можно задать с точностью до 0,01 °С. Пусть градации имеют вид: 0 – 3; 3 – 6; 6 – 9; 9 – 12.

После рекомендуемой операции они будут выглядеть так: 0,01 – 3,01; 3,01-6,01; 6,01 – 9,01; 9,01 – 12,01.

Вопрос о том, в какую градацию отнести значение температуры в 3 °С, снимается.

Вторым вариантом решения поставленного вопроса может служить следующий прием. После определения границ градаций принимается условие, что при совпадении значений случайной величины с границей градации, оно относится, например, только к левой из смежных градаций. В этом случае границы градаций можно выражать наиболее удобными числами, включая и целые. После определения количества градаций и их ширины строится ряд (Табл. 2.4) дифференциального распределения метеорологических элементов.

Таблица 2.4

Дифференциальное распределение случайной величины $\hat{x}$.

Численность	Границы градаций					
	$x_1 - x_2$	$x_2 - x_3$	…	$x_i - x_{i+1}$	…	$x_k - x_{k+1}$
m_i	m_1	m_2	…	m_i	…	m_k
p_i*	p_1*	p_2*	…	p_i*	…	p_k*
w_i	w_1	w_2	…	w_i	…	w_k
$w_{i\text{отн}}$	$w_{1\text{отн}}$	$w_{2\text{отн}}$	…	$w_{i\text{отн}}$	…	$w_{k\text{отн}}$

В этой таблице представляются численности статистического распределения, под которыми понимают величины, выражающие в каждой из градаций.

Численностями дифференциального статистического распределения служат: абсолютная частота (просто частота), относительная частота (частотность), абсолютная плотность, относительная плотность градаций.

Абсолютной частотой (частотной) градаций называется число случаев попадания наблюденных значений случайной величины (элементов выборки) в данную градацию. Она показывает, как часто могут появляться значения случайной величины из данной градации. Абсолютную частоту i-ой градации будем обозначать через m_1. Очевидно, что $\sum_{i=1}^{k} m_i$ = n, где n – объем архивной выборки.

Абсолютная частота обладает тем недостатком, что она не является показательной характеристикой для выборок разных объемов. Если требуется проводить сравнительный анализ статистических распределений для выборок разных объемов, то удобнее пользоваться понятием относительной частоты.

Относительной частотой (частотностью) градаций называется отношение абсолютной частоты соответствующей градации к общему числу обрабатываемых наблюдений n. Обозначается относительная частота обычно через p_1* и вычисляется по формуле:

$$p_1* = \frac{m_1}{n}. \qquad (2.5)$$

Из определения следует, что $0 \leq p_1* \leq 1$, и сумма частот всех градаций равна единице, т. е. $\sum_{n=1}^{k} p_1 * = 1$.

Относительная частота градаций является оценкой вероятности попадания случайной величины в данную градацию. В метеорологии относительную частоту обычно называют повторяемость.

Абсолютная и относительная частоты являются основными численностями дифференциального распределения, и в случае, если все градации равны между собой, ограничиваются их вычислениями. Если же в статистическом распределении имеются градации равной ширины, то эти численности теряют свою наглядность и показательность. Тогда рассчитываются абсолютные и относительные плотности.

Абсолютной плотностью градации называется отношение абсолютной частоты данной градации к ее ширине Δx, т. е.

$$w_i = \frac{m_i}{\Delta x_i}, \qquad (2.6)$$

Относительной плотностью градации называется отношение относительной частоты данной градации к ее ширине, т. е.

$$w_{i\text{отн}} = \frac{p_i*}{\Delta x_i}, \qquad (2.7)$$

Относительная плотность градации является приближенной оценкой плотности распределения вероятностей в данном интервале для серединного значения i-ой градации, т. е. $w_{i\text{отн}} = f(x_1)$.

Отсюда можно сделать вывод, что статистический ряд, численностями которого являются относительные плотности градаций, может служить подходящей параметрической оценкой (статистическим анализом) плотности распределения исследуемой случайной величины. При этом, чем больше будет объем обрабатываемой выборки, чем меньше будут градации, тем точнее статистический ряд будет воспроизводить теоретическую плотность распределения исследуемой случайной величины.

Для получения интегральных характеристик статистических распределений строят таблицы вида 2.5.

Таблица 2.5

Интегральное распределение случайной величины $\hat{x}$.

Численность	Границы градаций					
	x_1	x_2	…	x_i	…	x_k
m(X< x_1)	m_1 (X< x_1)	m_2 (X< x_2)	…	m_i (X< x_i)	…	m_k (X< x_i)
p*(X< x_1)	p_1* (X< x_1)	p_2* (X< x_2)	…	p_i* (X< x_i)	…	p_k* (X< x_k)

Здесь $x_1, x_2, \ldots, x_k$ – правые границы градаций.

В интегральных распределениях обычно представляют накопленные численности, причем накопление идет от меньших величин к большим. Иногда рассматриваются накопленные численности, когда накопление идет от больших значений случайной величины к меньшим. В этом случае $x_1, x_2, \ldots, x_k$ будут являться левыми гранями градаций, а численности иметь вид: m(X> x_1), p*(X> x_1). Рассмотрим случаи интегрального статистического распределения, когда накопление идет от меньшего значения случайной величины $\hat{x}$ к большему.

Абсолютной накопительной частотой называется число случаев попадания наблюдений значений случайной величины $\hat{x}$ (элементов выборки) в интервал от -∞ до правой границы i-ой градации x_i, т. е. это есть число случаев наблюдений значений величины $\hat{x}$ меньших, чем x_i. Абсолютная накопительная частота обозначается через m(X< x_i) и вычисляется обычно с использованием дифференциального статистического ряда по формуле:

$$m(\hat{X} < x_i) = \sum_j^i m_j, \qquad (2.8)$$

где x_i – правая граница i-ой градации;

m_j – абсолютная частота j-ой градации;

$m(X < x_i)$ – есть сумма частот от 1-ой до i-ой градации.

Относительной накопленной частотой называется относительная частота наблюдений случайной величины $\widehat{X}$ в интервале от - ∞ до x_i. Она обозначается через $P^*(\widehat{X} < x_i)$ и вычисляется обычно с использованием данных дифференциального статистического распределения по формуле:

$$P^*(\widehat{X} < x_i) = \sum_j^i p_j^*, \qquad (2.9)$$

где x_i – правая граница i-ой градации;

p_j^* – относительная частота i-ой градации;

Накопленные частоты иногда еще называют кумулятивными частотами.

Из приведенных определений следует, что статистический ряд, численностями которого являются накопленные частоты, может служить подходящей оценкой функции распределения $\widehat{X}$. Такую оценку функции распределения F(x) принято называть статистической функцией распределения и обозначать через F*(x). Иногда она называется выборочной функцией распределения или эмпирической функцией распределения и имеет вид:

$$p^*(x_i) = p^*(X < x_i) = \sum_j^i p_j^*, \qquad (2.10)$$

Формально статистическая функция распределения F*(x) обладает всеми свойствами истинной функции распределения случайной величины $\widehat{X}$, т. е. F(x). Однако ее значения дают не вероятности, а частоты неравенства $X < x$ в данной выборке. По теореме Бернули из закона больших числе при увеличении объема выборки значения $F^*(x_i)$ при каждом x_i сходятся по вероятности к $F(x_i)$.

Если для практических целей требуется знать абсолютную или относительную частоту появления значений $\widehat{X}$, больших некоторого заданного x_i, то используются формулы:

$$m(\widehat{X} > x_i) = \sum_j^k m_j, \qquad (2.11)$$

$$P^*(\widehat{X} > x_i) = \sum_j^k p_j^*, \qquad (2.12)$$

где x_i – левая граница i-той градации.

Рассмотрим пример обработки метеорологического ряда наблюдений в целях изучения закона распределения исследуемой величины.

Пример 3. Составить по данным таблицы месячных сумм осадков статистические ряды дифференциального и интегрального распределения январских сумм осадков по станции Воронеж.

Таблица 2.6

Месячные суммы январских осадков по станции Воронеж.

Год	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958
Сумма осадков, мм	36,5	5,2	41,2	34,4	16,6	15,0	62,2	32,2	15,5	23,8
Год	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968
Сумма осадков, мм	52,1	55,1	14,5	36,8	27,1	15,6	15,1	88,6	54,5	95,3
Год	1969	1970	1971	1972	1973	1974	1975	1976	-	-
Сумма осадков, мм	9,2	65,8	17,5	11,7	25,3	11,4	42,8	35,3	-	-

Из таблицы 2.6 находим x_{min} = 5.2 мм и x_{max} = 95,3 мм.

По формуле (2.1) (т. к. распределение одномерное) находим размах выборки:

$$L = 95,3 – 5,2 = 90,1 \text{ мм}$$

С использованием неравенства (2.2) определяем число градаций, на которое следует разбить значение случайной величины (январские суммы осадков):

$$k \leq 5 \lg n = 5 \lg 28 \; b = 7.2$$

Первоначально принимаем k = 7.

По формуле (2.4) находим ширину градаций

$$\Delta X = \frac{L}{k} = \frac{90.1}{7} \approx 12.3.$$

Задаемся целью сделать границы градаций целыми и ширину градаций удобной для анализа. Примем, что $\Delta x = 15$. В качестве левой границы 1-ой градации возьмем $x_1 = 0$. Тогда седьмая градация будет иметь левой границей значение равное 90, а правой – 105. Эти значения нас устраивают, и принимаем окончательно, что k = 7, $\Delta x = 15$, $x_1 = 0$. Принимаем, что если значение случайной величины совпадает со значением границы двух смежных градаций, то этот случай относится к левой смежной градации.

Составляем табл. 2.7 дифференциального статистического распределения, определив по данным табл. 2.7 и формулам (2.5, (2.6) и (2.7) численности соответствующих градаций.

Для проверки правильности расчетов необходимо понимать, что в первой строке сумма абсолютных частот должна быть равна объему архивной выборки, а сумма относительных частот равна 1.

Таблица 2.7

Дифференциальное распределение январских месячных сумм осадков по станции Воронеж

численность	Границы градаций, мм						
	0 – 15	15 – 30	30 – 45	45 – 60	60 – 75	75 – 90	90 – 100
mi	6	8	7	3	2	1	1
p_i^*	0.214	0.286	0.25	0.107	0.071	0.036	0.036
Wi	0.4	0.53	0.47	0.2	0.13	0.07	0.07
$w_{i\text{отн}}$	0.0143	0.019	0.0167	0.007	0.0046	0.0024	0.0024

Составляем табл. 2.8 интегрального статистического распределения, вычислив накопленные частоты.

Таблица 2.8

Интегральное распределение январских месячных сумм осадков по станции Воронеж

Численность	Границы градаций, мм						
	15	30	45	60	75	90	105
m(X< x_i)	6	14	21		26	27	28
p*(X < x_i)	0,214	0,5	0,75	0,857	0,928	0,964	1,0

2.3 Графическое представление статистических распределений.

Для большей наглядности, статистический ряд распределения случайной величины представляется графически в виде гистограммы, полигона, огивы. Часто график облегчает выбор подходящего вида функции для аналитического описания распределения анализируемого ряда наблюдений.

Гистограмма – один из видов представления статистического распределения случайной величины. Гистограмма дифференциального распределения строится следующим образом. Но оси абсцисс выбранной прямоугольной системы координат отмечаются точки $x_1, x_2, \ldots, x_{k+1}$, соответствующие границам градаций представляемого статистического ряда распределения. Затем на каждом из отрезков $[x_i\ x_{i+1}]$, i = 1, 2, …, k, как на основании, строится прямоугольник с высотой, равной численности соответствующей градации (абсолютной частоте, повторяемости и т. д.). На рис. 2.1 и 2.2 представлены гистограммы распределения абсолютных и относительных частот.

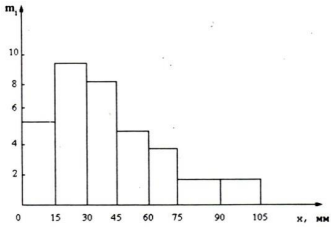


Рис. 2.1 Гистограмма распределения абсолютных частот

Гистограмма интегрального распределения (накопленных численностей) строится аналогично гистограмме дифференциального распределения. На оси абсцисс выбранной прямоугольной системы координат отмечаются границы градаций статистического ряда распределения x_1, x_2, …x_{k+1}. Затем на каждом из полученных отрезков $[x_i\ x_{i+1}]$, как на основании, строится прямоугольник, высота которого равна сумме частот всех градаций, лежащих левее правой границы градации, для случая накопления от меньших значений к большим, и правее левой границы, для случая накопления от больших значений случайной величины $\widehat{X}$ к меньшим. После этого верхние стороны построенных прямоугольников выделяются жирными линиями. В результате такой операции получается разрывная ступенчатая кривая, которая представляет собой график статистической функции распределения. Точки разрыва на графике накопленных частот соответствуют границам градаций. На рис. 2.3 показана ступенчатая кривая распределения накопленных частот месячных сумм осадков, построенная по данным примера 3.

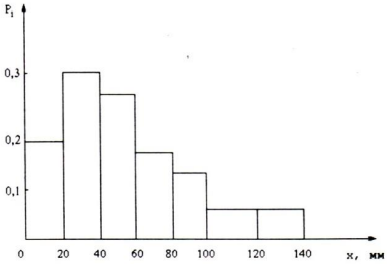


Рис. 2.2 Гистограмма распределения относительных частот

Кроме гистограмм для графического отображения статистических распределений используются ломаные линии, называемые полигоном или огивой.

Полигон представляет собой график, на котором распределение частот по градациям изображается точками, соединенными последовательно отрезками прямых. Он строится следующим образом. На оси абсцисс выбранной прямоугольной системы координат отмечаются середины градаций представляемого

статистического ряда. Затем на график наносятся точки, абсциссами которых являются середины градаций, а ординатами – соответствующие им частоты, и соединяются последовательно отрезками прямых.

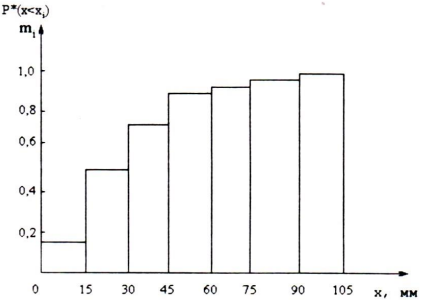


Рис. 2.3 Гистограмма интегрального распределения

Полигон можно получить и из гистограммы дифференциального распределения, если соединить отрезками прямых середины верхних сторон прямоугольников. На рис. 2.4 показан полигон распределения относительных частот месячных сумм осадков, построенный по данным примера 3.

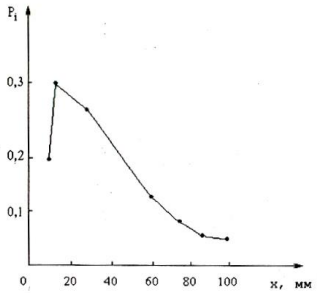


Рис. 2.4 Полигон распределения

Огива, (интегральная кривая распределения) является способом графического представления накопленных численностей. Для ее построения на поле графика наносятся точки, соответствующие границам градаций, а ординатами – соответствующие им накопленные частоты. Затем эти точки соединяются отрезками прямых.

Аналогично полигону огиву можно построить по гистограмме интегрального распределения. Для этого нужно соединить отрезками прямых правые вершины прямоугольников. Огива наиболее выразительно представляет вид статистической функции распределения случайной величины (рис. 2.5).

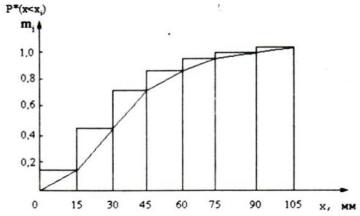


Рис. 2.5 Огива (интегральная кривая распределения)

Кумулята. Гистограмма интегрального распределения по своей сути является выборочной функцией распределения дискретной случайно величины. Для непрерывной случайной величины с помощью полигона или гистограммы можно построить так называемую кумуляту распределения путем интегрирования функции $\varphi(x)$.

$$F(x) = \int_{-\infty}^{x} \varphi(x)dx$$

График функции F(x) представлен на рис. 2.6. Отметим, что для дискретной случайной величины кумуляты не существует.

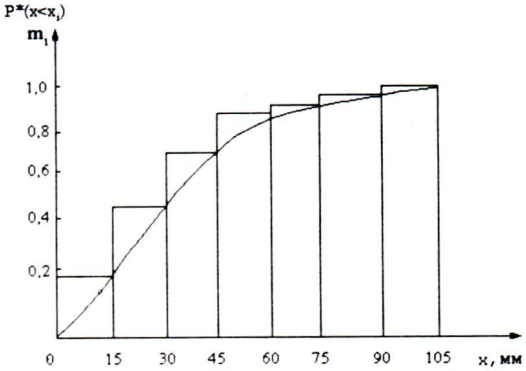


Рис. 2.6 Кумулята распределения

3 ОСНОВЫ ТЕОРИИ СТАТИСТИЧЕСКИХ РЕШЕНИЙ

3.1 Задачи принятия статистических решений

Все понятия теории вероятностей и математической статистики чрезвычайно полезны синоптику, как при метеорологическом обеспечении конкретного потребителя, так и при изучении процессов, происходящих в атмосфере, при уточнении существующих или разработке новых способов прогноза погоды.

Ответы на вопросы такого рода как - будет ли обостряться атмосферный фронт, какая будет высота нижней границы облачности и дальность видимости, возникает ли на аэродроме гроза, являются необходимыми этапами метеорологического обеспечения, а на вопросы выписывать или нет штормовое предупреждение, будут ли соответствовать метеорологические условия возможностям выполнить поставленную задачу и другие, являются ли конечной целью деятельности потребителя. Для ответа на указанные вопросы необходимо сделать некоторые заключения о состоянии атмосферы, о возможностях прогностического метода и условиях его использования, т. е. принять решение.

Решением называется некоторое заключение, вывод об исследуемом объекте или его свойствах.

Особенностью деятельности метеоролога является то, что он постоянно принимает решения в условиях неопределенности. Эта неопределенность постоянно изменяется, она может уменьшаться или увеличиваться, но всегда присутствует. Связана она с неполнотой информации о состоянии атмосферы, несовершенством прогностических моделей, различного рода помехам. В связи с этим информация носит статистический характер и принимаемые на ее основе решения являются статистическими.

Статистическим решением называется некоторое заключение (вывод) об исследуемом объекте или его свойствах, получено в результате статистической обработки информации.

В общем случае задача принятия статистических решений формулируется следующим образом. Имеется некоторый объект наблюдения, который может находиться в одном из m состояний. В процессе наблюдений за объектом получена некоторая совокупность результатов наблюдений $\hat{x}_1$, $\hat{x}_2$, …, $\hat{x}_n$, характеризующих состояние объекта в момент наблюдения. Необходимо на основе обработки данной совокупности результатов наблюдения принять некоторые решения о состоянии, в котором находится исследуемый объект в данный момент времени или в котором он будет находиться через некоторый промежуток времени.

Легко видеть, что в первом случае речь идет о диагнозе, а во втором – о прогнозе.

Пример 1. Известно, что циклоны в процессе своего развития 4 стадии (m = 4), причем каждой стадии соответствует определенный вид термобарического поля, определенное положение атмосферных фронтов относительно центра, глубина циклона, наклон пространственной оси и т. д. Синоптик получает комплект синоптических карт и карт барической топографии, по которому можно определить каждую из указанных характеристик ($\hat{x}_1$, $\hat{x}_2$, …, $\hat{x}_{n1}$). Требуется на основе анализа этих характеристик указать, в какой стадии развития находится циклон.

Любое принятие решения сопровождается получением некоторого «выигрыша» или некоторых «потерь». Пусть, например, синоптик разработал прогноз, в котором указал наличие интенсивных осадков. Если колхозники, следуя этому прогнозу, не будут поливать поля и осадки действительно пройдут, то выигрыш будет заключаться в экономии сил и средств, используемых для полива. Если же осадков не будет, то урожай может засохнуть и колхоз получит определенные потери.

Потерями называются отрицательные последствия, сопровождающие реализацию принятого решения. Виды потерь могут быть различными: расходы, связанные с реализацией решения; проигрыш при игре в карты; ущерб, полученный потребителем и т. д.

Полезностью называются положительные последствия, сопровождающие реализацию принятого решения.

Очевидно, что для принятия решения необходимо задать некоторый алгоритм, позволяющий связать множество результатов наблюдений с множеством возможных решений. Правило, устанавливающее закон выбора решения на множестве совокупностей сведений об изучаемом объекте, называется решающим правилом или решающей функцией. Существует много решающих правил, которые могут быть использованы при осуществлении того или иного вида обработки информации. Очевидно, что из этих правил следует применять то, которое обеспечивает принятие решения требуемого качества, т. е. решения оптимального в определенном смысле. В связи с этим возникает задача определения оптимального решающего правила.

Выбор такого правила определяется рядом факторов:

- требованиями, которые предъявляются к качеству решения;
- свойствами совокупности результатов наблюдений;
- условиями, в которых получаются результаты наблюдений;
- дополнительной априорной информацией, которая может быть использована при принятии решения.

Требования к качеству решения определяются потребителем решения. Так, например, если разрабатывается прогноз высоты облаков, то метеорологу нужно,

чтобы этот прогноз давал минимальное отклонение рассчитанных значений от фактических, а летчику, имеющему минимум

100 * 1, нужно точно знать, будет нижний край облачности выше 100 м или ниже. Причем, летчика не интересуют величины ошибок в прогнозе высоты нижнего края облачности для высот более 100 м. Если в прогнозе указывалось, что высота нижнего края облачности будет 300 м, а на самом деле ее высота оказалась 600 м, то для метеоролога это очень большая ошибка и прогноз принимается плохим. Для летчика же это хороший прогноз, так как он позволил ему выполнить полет. Если же по прогнозу высота нижнего края облачности ожидалась 110 м, а фактически оказалась 90 м, то для метеоролога такой прогноз будет хорошим (ошибся всего на 20 м), а для летчика – плохой, так как при 110 м он мог осуществить полет, а при 90 м – нет.

В данном примере мы имеем дело с двумя решениями. Первое – это решение метеоролога на формулировку прогноза, а второе – решение летчика на выполнение полета.

Условия, в которых получаются результаты наблюдений, можно разделить на две группы: условия пассивного эксперимента и условия активного эксперимента. В первом случае не осуществляется специальной организации (планирования) эксперимента с целью получения результатов наблюдения с необходимыми свойствами. Во втором случае эксперимент организуется так, чтобы полученная информация обладала всеми необходимыми свойствами: точностью, надежностью и др.

Свойства совокупности результатов наблюдений существенно влияют на качество решающей функции. К ним относятся:

- объем совокупности результатов наблюдений;
- представительность результатов;
- статистическая устойчивость результатов;
- однородность результатов;
- отсутствие в результатах грубых ошибок.

Влияние данных свойств на качество решающих функций будет рассмотрено в последующих разделах учебника.

Наиболее важными свойствами априорной информации являются ее объем и достоверность.

Основные элементы задачи принятия решений можно формализовать следующим образом:

задается множество А, называемое пространством действий, которое состоит из всех действий а ϵ А, доступных лицу принимающему решение;

задается множество М, называемое пространством параметров, состоящее

из всех возможных «состояний природы», m ϵ M, из которых может реализоваться одно и только одно, причем это истинное состояние природы неизвестно в момент, когда нужно принять решение;

задается функция L, называемая функцией потерь, которая каждой паре всех сочетаний возможных действий и «состояний природы» ставит в соответствие некоторое число, определяющее возможные потери; пара (m, a) называется последствием от принятия решения a, если истинное состояние природы есть m;

наблюдается случайная величина, $\widehat{X}$ возможные реализации которой

x ∈ $\widehat{X}$ образуют выборочное пространство $\widehat{X}$ и распределение которой задается плотностью распределения вероятностей f(x/m);

определяется множество D, называемое пространством решений и состоящее из всех отображений d множества $\widehat{X}$ в A.

Интерпретируем этот формализм следующим образом. В момент выбора действий лицо, принимающее решение, не знает истинное состояние природы и поэтому ему неизвестны истинные последствия его действий. Однако оно знает потери при каждом возможном последствия (m, a), определяемом его выбором действия a ∈ A и состоянием природы m ∈ M. Чтобы уменьшить неопределенность относительно m, лицо, принимающее решение, получает информацию в виде наблюдений случайной величины $\widehat{X}$, распределение которой зависит от состояния природы m. зная, что $\widehat{X}$ = x, и зная вид f(x/m), можно извлечь информацию относительно состояния m, которая может помочь в выборе общей стратегии, определяющей выбор a для каждого x. Формально действие выбирается на основе имеющегося наблюдения x. Выбор общей стратегии, определяющей для каждого x действие a, эквивалентен выбору решающей функции d ∈ D. Функция d определяет действие, предпринимаемое лицом, принимающим решение при всех возможных x = $\widehat{X}$.

Таким образом, теорию принятия решений можно считать наукой о том, как выбирать решающую функцию d из пространства решений D. Эта теория решает две принципиально важные задачи: первая состоит в выборе критерия для сравнения элементов множества D, вторая – в нахождении элемента d, оптимального в смысле выбранного критерия.

Иногда задачу принятия решения рассматривают как игру против природы. Это естественно, поскольку природа выбирает элемент m ∈ M, и затем лицо, принимающее решение, не зная этого состояния m, выбирает действие a. в результате этих двух выборов лицо, принимающее решение, терпит потери L(m/a), которые могут измеряться как в рублях, так и в потерях времени, затратах горючего и т. д

. Возможность наблюдать случайную величину $\hat{X}$ с плотностью распределения вероятности f(x/m) дает возможность использовать некоторую ограниченную информацию относительно выбора природы. Выбор решающей функции d можно рассматривать как стратегию, избранную для этой игры. Понятие стратегии является очень важным в теории статистических решений.

Стратегией лица принимающего решение (потребителя метеоинформации) будем называть совокупность логических или математических правил, предписывающих потребителю определенную линию поведения в каждой конкретной ситуации. Из этой формулировки следует, что стратегия представляет из себя не единичное действие, а некоторый общий принцип, которым потребитель должен руководствоваться постоянно, во всех случаях.

Отметим, что различают два вида стратегий: чистую и смешанную (рандомизированную). Стратегия называется чистой, если в каждом конкретном исходном состоянии природы выбирается одно и то же действие. Стратегия называется смешанной (рандомизированной), если каждой ситуации задается соответствующее распределение вероятностей различных действий.

В принципе, теория статистических решений является наиболее общим и фундаментальным разделом математической статистики. Все последующие разделы курса можно рассматривать как частный случай теории статистических решений. Сюда относится и теория оценивания, и теория проверки гипотез, и различные виды статистического анализа.

3.2 Показатели и критерии

Для математического описания системы атмосфера – метеорологическая информация потребителю необходимо дать детальное описание множества решений, которые могут приниматься в зависимости от складывающихся метеорологических условий и множеств состояний погоды, изменение которых влияет на процесс хозяйственной деятельности.

В зависимости от особенностей рассматриваемой задачи множества A (пространство действий) может быть дискретным или непрерывным. В соответствии с этим в первом случае должны быть перечислены все возможные действия потребителя a_1, a_2, …a_n, а во втором – указана некоторая область допустимых хозяйственных решений. Если речь идет о прогнозе опасных явлений для авиации явлений погоды: грозы, тумана, высоты облачности ниже установленного минимума, то, как правило, рассматривается только два варианта действий: летать (a_1) или не летать (a_2). Если же речь идет о работе котельной, отпуск тепла которой зависит от температуры наружного воздуха, то процесс можно

рассматривать как непрерывный. В этом случае в зависимости от прогноза температуры, принимается решение на изменение подачи тепла.

При разработке системы гидрометеорологического обеспечения следует учитывать реальные возможности использования соответствующей метеорологической информации и, наоборот, при разработке стратегии своей деятельности потребитель должен принимать во внимание специфику существующей системы гидрометеорологического обеспечения.

Таким образом, описание множества состояний погоды. Как и описание множества решений потребителя, представляет собой комплексную задачу.

Функция потерь

$$L = L(m, a), \qquad (3.1)$$

Показывает, каким будет экономический эффект в результате некоторого действия $a \in A$ при осуществлении погодных условий $m \in M$. эта функция часто имеет и другие названия – функция ущерба, функция потерь и полезности, функция выигрыша, платежная функция (матрица) и т. п. Несмотря на отсутствие здесь единой терминологии, во всех случаях речь идет об одной и той же количественной зависимости, характеризующей практическую значимость результатов различных действий во всевозможных метеорологических ситуациях.

Форма представления функции потерь (выигрышей) непосредственно зависит от того, в каком виде задаются множества D и M. Чаще всего встречаются случаи, когда эти множества дискретны и функции L = L(m, d) записывается в виде прямоугольной матрицы, содержащей число строк, равное числу возможных состояний погоды k, и число столбцов, равное числу S возможных хозяйственных решений. Общий вид такой матрицы представлен в табл. 3.1.

В реальных задачах числа k и S невелики. Наибольший интерес с точки зрения типичности встречающихся ситуаций при принятии решений представляет собой простейшая альтернативная схема, когда k = r = 2. Соответствующая матрица потерь представлена в табл. 3.2.

Таблица 3.1

Общий вид матрицы потерь (выигрышей) при произвольных k и S

M	D					
	d_1	d_2	…	d_j	…	d_s
m_1	l_{11}	l_{12}	…	l_{1j}	…	l_{1s}
m_2	l_{21}	l_{22}	…	l_{2j}	…	l_{2s}
…	…	…	…	…	…	…
m_i	l_{i1}	l_{i2}	…	l_{ij}	…	l_{is}
…	….	…	…	…	…	…
m_k	l_{k1}	l_{k2}	…	l_{kj}	…	l_{ks}

Таблица 3.2

Матрицы потерь 2 х 2

M	A	
	d_1	d_2
m_1	l_{11}	l_{12}
m_2	l_{21}	l_{22}

Такие матрицы обычно используют, когда речь идет о прогнозе опасного явления погоды, которое может произойти (состояние m_1) или не произойти (состояние m_2), а в зависимости от ожидаемых условий принимается одно из двух возможных решений. Пусть, например, в жаркий летний день по прогнозу ожидается возникновение грозы, а продавщица мороженного может осуществить торговлю либо в магазине (решение d_1), либо на пляже (решение d_2).

В том случае, если продавщица поверит прогнозу, останется в магазине и гроза возникнет (состояние m_1), народ побежит прятаться от грозы в магазине и раскупит мороженное. При этом продавщица получит прибыль l_{11}. Если она прогнозу в этой ситуации не поверит и выедет на пляж, то людей на пляже не будет и она сама промокнет. Ее убытки будут равны l_{12}. Если же по прогнозу ожидается отсутствие грозы, продавщица выехала на пляж и грозы не было, то прибыль ее будет равна l_{22}. Если останется в магазине, то убытки будут l_{21}.

Очевидно, что любые действия целесообразны только тогда, когда их стоимость ниже возможного ущерба от неблагоприятной погоды.

Во многих случаях при построении метеоролого-экономических моделей удобнее оперировать не самими значениями потерь L(m, d), а некоторыми связанными с ними разностными величинами R(m, d), показывающими, в какой степени увеличиваются потери или, соответственно, снижается прибыль, когда вместо оптимального действия выбирается другое. Если предположить, что исходные величины l(m, d) имеют смысл потерь, то функция R(m, d),задаваемая на всевозможных парах (m, d), будет записываться как

$$R(m, d) = L(m, d) - L(m, d) \quad (3.2)$$

где

$$L(m, d_0) = \min_{d \in 0} l(m, d) \quad (3.3)$$

Аналогично, если l(m, d) – выигрыши, то

$$R(m, d) = L(m, d_0) - L(m, d), \quad (3.4)$$

где

$$L(m, d_0) = \max_{d \in 0} l(m, d) \quad (3.5)$$

Величины R в обоих случаях положительны и представляют собой потери, которые обусловлены несоответствием выбранных действий фактическим погодным условиям. Функцию R(m, d) принято называть риском (сожалением).

Исходя из того, что по своему смыслу величины R характеризуют убытки, которые возникают из-за непредусмотрительного отклонения фактических погодных условий от предполагавшихся, положенных в основу принятия решения, такие потери называют метеорологическими.

Рассмотрим процедуру перехода от величины l к R.

Пример 2. В табл. 3.3 указаны значения потерь l_{ij}, получаемых некоторым потребителем метеорологической информации при различных сочетаниях фаз погоды m_i и решений d_j.

Таблица 3.3

Матрица потерь l

M	D		
	d_1	d_2	d_3
m_1	43	24	31
m_2	90	42	30
m_3	51	68	70

Согласно формулам (3.2) и (3.3), в данном случае для перехода к метеорологическим потерям в каждой строке исходной матрицы потерь надо найти минимальный элемент min l(m, d) и произвести построчную замену величин l_{ij} разностями

$$R\ (m_i, d_j) = l\ (m_i, d_j) - \min l\ (m_i, d_j), \qquad (3.6)$$

В соответствии с указанным преобразованием, например, для первой строки получим R (m_1, d_1) = 19, R (m_1, d_2) = 0, R (m_1, d_3) = 7.

Найденная таким образом матрица метеорологических потерь представлена в табл. 3.4.

Таблица 3.4

Матрица метеорологических потерь

M	D		
	d_1	d_2	d_3
m_1	19	0	7
m_2	60	12	0
m_3	0	17	19

Одним из возможных путей оценки качества решающей функции d в показателях, которые можно вычислить, является определение того, насколько хороша выбранная стратегия «в среднем». Для этого используется такое понятие, как функция риска.

Функцией риска называется закон, ставящий в соответствие принятому решению d и состоянию погоды m некоторые средние потери. Определяется функция риска равенством

$$R(m, d) = \int_x L(m, d(x)) f(x/m) dx \quad (3.7)$$

для непрерывных случайных величин или

$$R(m, d) = \sum_{x \in \hat{X}} L\left(m, d(x)\right) p(x/m) \quad (3.8)$$

для дискретных случайных величин.

Функция риска R(m, d) имеет смысл меры ожидаемых потерь от использования решающей функции d, если погода находится в состоянии m (математическое ожидание вычисляется по отношению к распределению, задаваемому функцией $f(x/m)$).

Задача статистического решения сводится к тому, чтобы выбрать такое решающее правило, которое обеспечило бы минимум риска или функции риска. Так, например, из данных табл. 3.4 следует, что при состоянии погоды m лучше принять решение d_2, при m_1 - d_3, при m_3 - d_1, так как они обеспечивают минимальный риск.

Видно, что минимизация риска в данном случае будет целью построения решающего правила. Поэтому часто такой критерий называют целевой функцией потребителя.

Из формул (3.7) и (3.8) следует, что мы некоторым функциональным распределениям L(m d) и $f(x/m)$ ставим в соответствие числовую меру R. Закон, по которому элементы некоторого множества ставятся в соответствие элементам числового множества, будем называть функционалом. Примером функционала является определенный интеграл, который функцию переводит в число. Неопределенный интеграл функционалом не является. С использованием этого понятия задачу выбора оптимального решающего правила можно сформулировать как требование минимизации функционала R(m, d).

В практической деятельности в качестве критериев или целевых функций кроме понятий риска и функции риска используют и другие величины. Выбор критериев определяется самой постановкой задачи. Если требуется выполнить задачу с максимальной вероятностью, то критерием, очевидно, должна являться максимальная вероятность выполнения задачи; если эту задачу необходимо выполнить в минимальное время, то критерий – минимальное время и т. д.

3.3 Байесовская стратегия

Предположим, что функция потерь, а соответственно и функция риска неизвестны, (на практике такие случаи встречаются достаточно часто). Решение на выполнение задачи вместе с тем принимать надо. В таких ситуациях используются правила, исключающие необходимость использования функции риска. Одним из таких правил является правило, основанное на принципе максимальной апостериорной вероятности. Рассмотрим суть этого правила на примере.

Пример 3. Пусть известно, что летчик может осуществить взлет в том случае, если высота облаков H будет больше некоторого значения H*, и взлет не может состояться при $H < H^*$. пусть, в свою очередь, высота облаков зависит только от одного фактора дефицита точки росы D. В результате измерения получилось некоторое значение дефицита D_i. Требуется определить, можно ли при этом значении дефицита D_i осуществить взлет.

Для решения такой задачи можно поступить следующим образом. Соберем все случаи, когда облачность была ниже установленного минимума H* и каждому случаю поставим в соответствие значение дефицита D_i. Разобьем эти случаи на соответствующие градации и для каждой градации определим вероятность того, что $H < H^*$. аналогичным образом построим закон распределения для ситуации $H > H^*$. в результате получим два закона распределения, представленных на рис. 3.1. если в результате измерения мы получим значение дефицита точки росы D_i, то, поднимаясь от этого значения до соответствующих кривых, можем получить значения вероятности того, что облачность будет выше установленного минимума, и вероятность того, что она будет ниже этого значения. Сравнивая полученные вероятности мы можем решить, какая будет облачность. Если $P(H > H^*) > P(H < H^*)$, то принимается решение, что высота облаков будет выше установленного минимума, в противном случае – наоборот. В соответствии с этим заключением принимается решение на полеты.

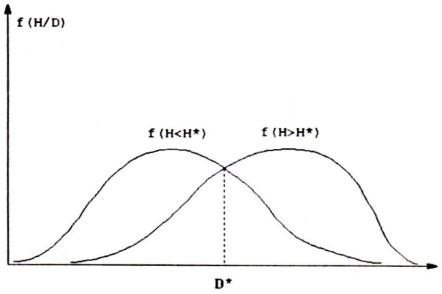


Рис. 3.1 Закон распределения D для случаев $H > H^*$ и $H < H^*$

В том случае, если имеются не две градации облачности, а четыре, то надо строить четыре закона распределения и выбирать в качестве решения тот, который при измеренном значении дефицита дает большую вероятность.

Аналитически эту задачу можно решить с использованием формулы Байеса (теоремы гипотез)

$$P(H_i/X_{<n>}) = \frac{p(H_i)f(X_{<n>}/H_i)}{f(X_{<n>})}, \qquad (3.9)$$

где H – решение (гипотеза);

$X_{<n>}$ – вектор исходных характеристик состояния атмосферы;

$f(\frac{X_{<n>}}{H_i})$ – условная плотность распределения X_n.

$$f(X_{<n>}) = \sum_{i=1}^{m} p(H_i) \quad f(X_{<n>}/ H_i) \qquad (3.10)$$

безусловная плотность распределения вектора $X_{<n>}$.

Сравнивая полученные $P(H_i/X_{<n>})$ выбираем то H_i, которое дает максимальное значение вероятности.

Следующим, часто используемым правилом принятия решения, является правило основанное на принципе минимальной вероятности ошибки.

Принцип минимальной вероятности ошибки применим в тех случаях, когда известны только условный закон распределения результатов наблюдений. Сущность данного принципа состоит в том, что минимизируется вероятность неправильного решения. Пусть в результате наблюдения получено некоторое значение x наблюдаемой случайной величины $\hat{X}$ и на основе его анализа необходимо принять решение о том. Что объект находится в состоянии m_i. Пусть объект имеет два возможных состояния m_1 и m_2 и кривые условных законов распределения $f(x/\hat{m}_1)$ и

$f(x / \hat{m}_2)$ имеют вид, представленный на рис. 3.2.

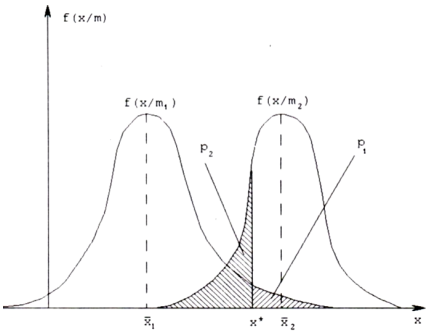


Рис. 3.2 Кривые условных законов распределения $f(x/\hat{m}_1)$ и $f(x / \hat{m}_2)$

Если принимать решение, основываясь на принципе максимальной апостериорной вероятности, то при $\hat{x} > x^*$ следует принять решение, что объект будет находиться в состоянии m_2, а при $\hat{x} < x^*$ - в состоянии m_1. Видно, что при систематическом принятии этих решений потребитель будет допускать ошибки, так как при $\hat{x} > x^*$ существует вероятность того, что объект окажется в состоянии m_1, а при $\hat{x} < x^*$ - в состоянии m_2. Эти ошибки носят специальные названия и заслуживают подробного рассмотрения.

При решении указанной задачи возможен ряд исходов, для описания которых введен следующие случайные события:

$\hat{A}$ – объект наблюдения находится в состоянии m_1;

$\hat{\bar{A}}$ – объект наблюдения находится в состоянии m_2;

$\hat{B}$ – принято решение о том. Что объект наблюдения находится в состоянии m_1;

$\hat{\bar{B}}$ - принято решение о том, что объект находится в состоянии m_2.

При принятии решения возможны следующие исходы:

$\hat{A} \cap \hat{B}$ - принято решение, что объект находится в состоянии m_2, и он действительно находится в этом состоянии;

$\hat{A} \cap \hat{\bar{B}}$ - принято решение том. Что объект находится в состоянии m_2, а он находится в состоянии m_1;

$\hat{\bar{A}} \cap \hat{B}$ - принято решение о том, что объект находится в состоянии m_1, а он находится в состоянии m_2;

$\hat{\bar{A}} \cap \hat{\bar{B}}$ - принято решение, что объект находится в состоянии m_2, и он действительно находится в этом состоянии.

Анализ этих исходов показывает, что в первом и последнем случае принимаются правильные решения, а во втором и третьем – ошибочные.

Ошибкой первого рода (ошибкой «пропуска») называется ошибка, представляющая собой принятие решения о том, что объект наблюдения не находится в предполагаемом состоянии (m_1), в то время, как в действительности он пребывает именно в этом состоянии (m_1).

Ошибкой второго рода (ошибкой «ложной тревоги») называется ошибка, представляющая собой принятие решения о том, что объект наблюдения находится в предполагаемом состоянии (m_1), в то время как в действительности он пребывает в другом состоянии (m_2). Очевидно, что в общем случае правило решения должно быть таки , чтобы обеспечивалось минимально возможная вероятность принятия ошибочных решений.

Для случая, представленного на рис. 3.2, при решении о том, что объект находится в состоянии $m_1(x < x^*)$, вероятность совершения ошибки будет равна

площади под кривой $f(\hat{x} / x_2)$ на участке (-∞, x*), которая описывается интегралом:

$$p_1 = \int_{-\infty}^{x^*} f(\hat{x} / m_2)dx \quad (3.11)$$

Если же будет принято решение о том, что объект находится в состоянии m_2, то вероятность ошибки p_2 будет определяться площадью под кривой $f(\hat{x} / m_1)$ на участке (x*, +∞)

$$p_2 = \int_{x^*}^{+\infty} f(\hat{x} / m_1)dx \quad (3.12)$$

Если последствия ошибочных решений оценить невозможно, то очевидно, что решающее правило должно обеспечить минимум суммы вероятностей p_1 и p_2, а так как это правило состоит в определении границы x*, то, следовательно, величина x* должна выбираться таким образом, чтобы минимизировать величину:

$$Y = \int_{-\infty}^{x^*} f(\hat{x} / m_2)dx + \int_{x^*}^{+\infty} f(\hat{x} / m_1)dx \quad (3.13)$$

Общих правил выбора оптимального значения величины x* не существует. Для симметричных законов распределения минимум суммы (3.13) достигает при выполнении условия

$$p_1 = p_2 \quad (3.14)$$

На практике очень часто стремятся минимизировать вероятность ошибки первого рода (попуска), устанавливая ее в пределах 0,01 – 0,1. На основании этой установки выбирается значение критической границы.

Ко второму классу задач, связанных с выбором оптимального решающего правила, относятся ситуации, когда функции риска являются известными. Из возможных вариантов решения таких задач рассмотрим только два подхода – минимаксный и байесовский.

Для объяснения сути этих подходов рассмотрим рис. 3.3, на котором изображены графики двух гипотетических функций риска, соответствующих решающим правилам d_2 и d_2 для некоторой задачи принятия решений с одномерным пространством состояний M.

Для значений M M_1 и M M_2 решающая функция d_2 приводит к меньшему риску, чем d_1, но для других M значение риска d_2 намного больше, чем при d_1.

В такой ситуации возможны два варианта принятия решения:

В первом варианте разумно проявить осторожность и выбирать d_1, так как при M M_1 и M M_2 R (m, d_1) незначительно больше, чем R (m, d_2), а при M_1 M M_2 R (m, d_1) намного меньше. Такое решение предохранит от наихудшего возможного исхода; оно минимизирует максимум возможного риска.

Второй вариант используется, если имеется дополнительная информация.

Тогда можно вынести предположение о том, каких значений M следует скорее всего ожидать. Если значения M окажутся в области M_1 M M_2, то очевидно, следует выбирать d_1. Если же M попадает в другие области, то разумно выбирать d_2.

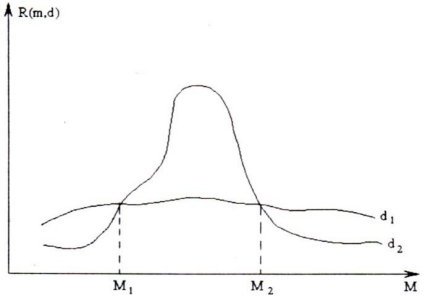


Рис. 3.3 Функция риска для решений d_1 и d_2

Эти два подхода легко формализовать.

В первом варианте выбираем такое решающее правило d*, чтобы

$$R(m, d^*) = \min_{d \in D} \max_{m \in M} R(m, d). \quad (3.15)$$

Другими словами, мы выбираем решающую функцию, для которой максимальное значение риска равно наименьшему возможному максимуму риска по всему пространству решений D. Поэтому d* называется минимаксной решающей функцией.

Во втором варианте предположение о параметре M может быть выражено в виде плотности распределения вероятностей $f(m)$ в пространстве M. функция риска R(m, d) выражает ожидаемые потери от использования решающей функции d при условии, что M – истинное состояние природы, которое нам заранее неизвестно. Если известно $f(m)$, то можно вычислить ожидаемое значение риска по отношению к предполагаемому распределению $f(m)$. Очевидно, что значение риска R(d), соответствующего решающей функции d, может быть определено с помощью выражения

$$R(d) = \int_M R(m, d) f(m) dm, \quad (3.16)$$

для непрерывного случая или

$$R(d) = \sum_{m \in M} R(m, d)\, p(m), \quad (3.17)$$

для дискретного случая.

Риск R(d) называется байесовским риском.

Естественным является желание выбрать решающую функцию, которая минимизирует средние ожидаемые потери. Назовем d* байесовской решающей функцией, если она удовлетворяет условию.

$$R(d^*) = \min_{m \in M} R(d). \quad (3.18)$$

Следует отметить, что в каждой задаче принятия решения функция d* не единственная, так как она зависит, в частности, от выбора распределения $f(m)$. Поэтому лучше говорить, что d* является байесовской решающей функцией по отношению к $f(m)$.

Мы рассмотрели случай двух решающих функций d_1 и d_2. При возможном увеличении количества этих функций принципиально суть задачи не меняется, но возникает дополнительные сложности. Для устранения этих сложностей разработаны различные методы, один из которых приводится далее.

Пример 4. Пусть имеется 4 возможных решений d_1, d_2, d_3, и d_4. Функции риска, соответствующие каждому решению, представлены на рис. 3.4. ясно, что решения d_1 и d для всех значений m имеют большие значения риска, чем d_3 и d_4. Поэтому эти решения можно сразу выбросить из дальнейшего рассмотрения и искать оптимальные решения только между d_2 и d_4. Формализовано эта задача выглядит так. Если для заданной решающей функции d' $\in$ D существует другая решающая функция d" $\in$ D, такая, что R(m, d") $\leq$ R(M, d') для любых m $\in$ M и R(m, d") $<$ R(m, d') для некоторого значения m $\in$ M, то говорят, что решение d' доминируется решением d" или, что решение d" доминирует решение d'.

Решающая функция доминируемая некоторой другой решающей функцией, называется недопустимой, а в противном случае – допустимой. На рис. 3.4 решающие функции d_1 и d_2 доминируются функциями d_3 и d_4, и поэтому являются недопустимыми. В классе D = (d_1, d_2, d_3, d_4) d_3 и d_4 являются допустимыми, поскольку ни одна из них не доминирует другую. Если взять большее множество D, но так, чтобы (d_1, d_2, d_3, d_4) было его подмножеством, то решающие функции d_3 и d_4 сами могут быть доминируемыми какими-нибудь другими функциями и поэтому недопустимыми. Допустимость всегда является относительным понятием, связанным с выбором конкретного множества D.

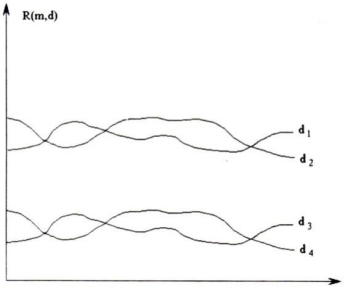


Рис. 3.4 Общая картина соотношений функции риска для нескольких решающих правил d

С точки зрения теории принятия решений с помощью концепции допустимости можно достичь гораздо большей ясности. Исключив недопустимые решения и сосредоточив усилия только на выборе допустимых решающих функций.

3.4 Нормированный риск и пороговая вероятность

Использование байесовского риска наиболее распространено в практике принятия статистических решений, однако, часто связано с определенными трудностями. Эти трудности возникают из-за незнания потребителем элементов матрицы потерь. Избежать этих трудностей можно с помощью введения понятия пороговой вероятности.

Пусть информации я о состоянии объекта регулярно используется для выбора одного из двух возможных планов деятельности d_1 и d_2, причем объект может находиться в одном из двух состояний m_1 и m_2. И пусть эффективность принятия решения определяется постоянной матрицей потерь (табл. 3.5).

Таблица 3.5

Постоянная матрица потерь

Состояние объекта	Решения	
	d_1	d_2
m_1	l_{11}	l_{12}
m_2	l_{21}	l_{22}

где l_{11} и l_{21} – потери потребителя, принявшего решение d_1, соответственно при состояниях объекта m_1 и m_2;

l_{12} и l_{22} – потери потребителя, принявшего решение d_2.

Если известна вероятность того, что объект окажется в состоянии m_1, и она равна p, то потери потребителя при выборе соответствующего плана действий определяется по формулам:

$$L(d_1) = p l_{11} + (1 - p)\, l_{21}, \quad (3.19)$$

$$L(d_2) = p l_{12} + (1 - p)\, l_{22}. \quad (3.20)$$

Сравнивая эти потери, можно выбрать такое решение, которое обеспечивает минимум потерь. Если $L(d_1) < L(d_2)$, то следует выбрать решение d_1, в противном случае - d_2.

Из сравнения формул (3.19) и (3.20) видно, что существует такая вероятность p*, при которой $L(d_1) = L(d_2)$, т. е. потери потребителя будут одинаковы при выборе обеих стратегий. Эту вероятность p* будем называть пороговой вероятностью. Значение пороговой вероятности может быть легко определено из условий равенства правых частей выражений (3.19) и (3.20)

$$p l_{11} + (1 - p)\, l_{21} = p l_{12} + (1 - p)\, l_{22}, \qquad (3.21)$$

Отсюда после несложных преобразований получим:

$$P^* = \frac{l_{22} - l_{21}}{p l_{11} - l_{21} + l_{22} - p l_{12}}. \qquad (3.22)$$

Так как p – это вероятность того, что объект будет находится в состоянии m_1, а 1 – p – в состоянии m_2, то естественно принять следующие решения. Если $p < p^*$, то следует готовиться к том, что объект будет находиться в состоянии m_2, если $p > p^*$ - в состоянии m_1 и принимать решения, которым эти состояния наиболее благоприятны. Таким образом, для рассматриваемого варианта задачи в рекомендациях достаточно указать способ определения знака разности p – p*.

Из этого следует, что пороговая вероятность объективная характеристика. Связанная с матрицей затрат. Если матрица неизвестна (что часто встречается в практике), то значение пороговой вероятности может быть получено с помощью метода экспертных оценок. Для этого специальным образом опрашиваются специалисты (эксперты), принимающие решение и отвечающие за его выполнение. Так, например, командирам авиационных полков задается вопрос: «При какой вероятности возникновения тумана вы отправите полк на запасный аэродром?» результаты этих опросов соответствующим образом обрабатываются и в результате получается значение пороговой вероятности p*.

Для потребителей с известной вероятностью легко определить критерий эффективности принятия решения. Предположим, что у нас имеется матрица затрат в виде табл. 3.5 и матрица сопряженности числа случаев принятых решений для соответствующих состояний объекта (3.6)

Таблица 3.6

Матрица сопряженности

Состояние объекта	Решения	
	d_1	d_2
m_1	n_{11}	n_{12}
m_2	n_{21}	n_{22}

Здесь n_{11} - число случаев в результате проведенных экспериментов, когда принималось решение d_1, а объект находился в состоянии

m_1 и т. д.

Суммарные потери потребителя по испытательной выборке из N случаев (N = $n_{11} + n_{12} + n_{21} + n_{22}$) составят

$$N = n_{11}l_{11} + n_{12}l_{12} + n_{21}l_{21} + n_{22}l_{22}. \quad (3.23)$$

В случае, когда состояние объекта предсказывается безошибочно (идеальный прогноз) при p > p* объект будет находится в состоянии m_1, а решение всегда будет d_1. При p <* - состояние m_2, решение - d_2, и суммарные затраты будут равны:

$$L_{ид} = (n_{11} + n_{12})\, l_{11} + (n_{21} + n_{22})l_{22}. \quad (3.24)$$

Очевидно, что чем меньше разность L - $L_{ид}$, тем более эффективной должна считаться методика принятия решения.

$$L - L_{ид} = n_{11}l_{11} + n_{12}l_{12} + n_{21}l_{21} + n_{22}l_{22} - (n_{11} + n_{12})\, l_{11} -$$

$$- (n_{21} + n_{22})l_{22} = n_{12}(l_{12} - l_{11}) + n_{21}(l_{21} - l_{22}).$$

Если нормировать эту разность на величину $l_{12} - l_{11}$, то получим:

$$\frac{L - L_{ид}}{l_{12} - l_{11}} = n_{12} + n_{21}\frac{l_{21} - l_{22}}{l_{12} - l_{11}}. \quad (3.25)$$

Легко показать, что

$$\frac{l_{21} - l_{22}}{l_{12} - l_{11}} = \frac{p*}{1 - p*}. \quad (3.26)$$

Обозначив левую часть через ф и учтя (3.26), выражение (3.25) преобразуем к виду

$$Ф = n_{12} + n_{21}\frac{p*}{1 - p*}. \quad (3.27)$$

Величину ф будем называть нормированным риском потребителя. Данный критерий удобно использовать при оценке эффективности способов прогноза опасных для авиации явлений погоды. Так при p* = 0,5 критерии ф становятся равными 1 – U, где U – оправдываемость прогнозов. Отношение p*/1-p* показывает во сколько раз ошибки пропуска становятся опаснее ошибок «ложной тревоги». Очевидно, что при выборе прогностического способа для обеспечения конкретной задачи лучшим должен признаваться тот, который обеспечивает минимум Ф.

3.5 Выбор стратегии при многокритериальных задачах

Метеорологическое обеспечение авиации имеет свою специфику по сравнению с другими видами метеорологического обеспечения. Прежде всего, эта

специфика заключается в использовании своеобразных критериев. Так, например, если эффективность метеорологического обеспечения сельского хозяйства или строительства можно оценить в денежных единицах, то для авиации такой подход далеко не всегда приемлем. Кроме экономической эффективности для авиации большую роль играет безопасность полетов. Причем критерии экономической эффективности и безопасности являются противоположными по смыслу. Для того. Чтобы обеспечить экономическую эффективность, авиация должна как можно больше летать, а для того, чтобы обеспечивать безопасность полетов – как можно больше находиться на аэродромах. Максимальная безопасность будет обеспечена тогда, когда ни один летательный аппарат не поднимется в воздух или когда полеты не будут зависеть от погодных условий. Поэтому метеорологическое обеспечение авиации должно сводиться к одновременному «удовлетворению» как минимум двух разных критериев.

С точки зрения метеообеспечения безопасность полетов может характеризоваться «степенью риска» R – вероятность возникновения предпосылок аварийной ситуации по метеоусловиям вследствие ошибочного прогноза. С экономической точки зрения эффективность использования метеоинформации определяется значением критерия экономической эффективности К, вид которого задается планирующими органами.

Таким образом, эффективность использования метеоинформации в рассматриваемом случае характеризуется двумя статистическими критериями – R и K, а задача выбора оптимального метода ее использования является типичной задачей теории многокритериальных методов принятия решений. Если множество допустимых методов использования метеоинформации не содержит метода, обеспечивающего одновременно необходимые экстремумы по обоим критериям (это возможно только при идеальном прогнозе соответствующей заблаговременности), для решения задачи необходимо в конечном счете установление некоторой иерархии между критериями.

В теории многокритериальных решений существует ряд методов установления межкритериальной иерархии, которые условно можно разделить на две группы: установление иерархии путем введения целевой функции на множестве критериев; установление допустимых сочетаний значений критериев с последующим экспертным отбором «лучших» сочетаний.

Рассмотрим ситуацию, когда требуется осуществлять метеорологическое обеспечение так, чтобы обеспечить соответствующим образом безопасность полетов. В этой ситуации основной задачей метеорологического обеспечения является построение таких прогностических методов, которые учитывают требова-

ния безопасности в рамках указанного подхода. Рядом авторов доказано, что такими прогностическими методами являются категорически вероятностные прогнозы. При категорически вероятностном прогнозе синоптик определяет не вероятность выполнения задачи по метеорологическим условиям P_B, а его соотношение с пороговой вероятностью P* (см. § 3.4). если $P_B > P^*$, то планируется выполнение задачи В, а если $P_B < P^*$, то не планируется. При этом несущественно, насколько P_B больше или меньше P*. Таким образом, синоптик должен лишь определить, больше P_B или меньше, чем P*, и соответственно рекомендовать выполнять или не выполнять задачу В. Прогнозист не должен выдавать потребителю значение P_B, так как любое решение, принятое потребителем, не может быть лучше предложенного прогнозистом, по определению являющегося оптимальным. Важно лишь, чтобы прогнозист рассматривал весь комплекс задач, стоящих перед потребителем.

Рассмотрим возможности построения методов категорическивероятностного прогноза, учитывающего требования безопасности полетов.

3.5.1 Метод критерия полной эффективности.

В явном виде влияние качества использования метеоинформации на экономическую эффективность деятельности потребителя проявляется в зависимости критерия экономической эффективности К от элементов матрицы сопряженности числа случаев принятых решений для соответствующих состояний объекта (табл.3.6) n_{ij}, где n_{ij} – вероятность принятия по прогнозу решения i, когда по реализовавшимся метеоусловиям следовало бы принять решение j. В качестве параметров К содержит экономические характеристики решаемых задач. Для большинства критериев такими параметрами являются элементы матрицы затрат l_{ij} (табл. 3.5), где l_{ij} – затраты при принятии по прогнозу решения i, когда по реализовавшимся метеоусловиям следовало бы принять решение j.

Очевидно, непосредственная оценка полных затрат при аварийных ситуациях невозможна. Однако, на основании экспертного опроса можно оценить «относительную стоимость» ошибок прогноза С, показывающую, во сколько раз ошибка пропуска аварийной ситуации «дороже» ошибки ложной тревоги. Более того, результаты испытаний показывают, что дисперсия оценки С для группы экспертов, относящихся к потребителям одного класса, достаточно мала. Таким образом, возникает возможность построения матрицы затрат L_0.

$$L_0 = \begin{pmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{pmatrix},$$

где l_{ij} - элементы матрицы экономических затрат, определяемые соответствующими экономическими службами.

На основании L_0 может быть построен критерий «полной» эффективности $К_0$, совпадающий по виду с К, но отличающийся от него значениями параметров, заменой элементов матрицы L на элементы матрицы L_0.

Для простейшей байесовской задачи, когда К – математическое ожидание затрат за фиксированный период времени, параметры l_{ij} однозначно определяют пороговую вероятность p* - минимальное значение вероятности не возникновения предпосылок аварийной ситуации, при которой экономически выгодно выполнять полет,

$$P_0^* = \frac{l_{21} - l_{11}}{l_{21} - l_{11} + l_{12} - l_{22}}. \quad (3.28)$$

Соотношение (3.28) позволяет, определив на основании экспертного опроса значение P*, учитывающее требования безопасности полетов, построить свою матрицу полных затрат L_0, все элементы которой должны совпадать с элементами матрицы L, за исключением

$$l_{12}^0 = l_{22} + \frac{1-P_0^*}{P_0^*}(l_{21} - l_{11}). \quad (3.29)$$

Предположение о существовании критерия полной эффективности $К_0$ справедливо, если совпадают элементы матриц полных затрат, построенных описанным способами. Нетривиальная связь между оцениваемыми экспертами значениями С и P_0^*, содержащая в качестве параметров элементы L, с одной стороны, делает такое несовпадение отнюдь не очевидным, а с другой стороны, при совпадении служит надежным подтверждением существования $К_0$.

$$C = \frac{1-P_0^*}{P_0^*}\left(1 - \frac{l_{11}}{l_{21}}\right) + \frac{l_{12}}{l_{21}}, \quad (3.30)$$

Если соотношение (3.30) выполняется, то существование однозначной функции полезности на множестве критериев для байесовского критерия К и критерия R можно считать доказанным. Однако необходимости в непосредственном построении этой функции нет, поскольку оптимизация ее значения эквивалента минимизация значения критерия полной эффективности $К_0$. Последняя задача решается методами категорически вероятностного прогноза, изложенными в учебник.

3.5.2 Метод допустимых значений R.

Реализация рассмотренного подхода требует от экспертов сопоставления возможных последствий аварий (в том числе гибели экипажа и пассажиров) с чисто экономическими факторами. Значительно более простым для экспертов является определение допустимой с точки зрения безопасности полетов (без учета конкретного экономического эффекта) степени риска R_m.

При заданном значении R_m категорически вероятностный прогноз строится следующим образом. Для каждой пары «прогностический метод – стратегия» из множества возможных определяется значение R, затем выделяется множество, на котором $R < R_m$, и на этом подмножестве выбирается пара, обеспечивающая оптимум К. иными словами. Определяется условный экстремум К при $R < R_m$.

3.5.3 Метод прямых экспертных оценок

Принципиальным моментом в описанных методах является необходимость предварительных экспертных оценок. Если таких нет, может быть использован метод прямых оценок, при котором эксперты или лицо, принимающее решение, непосредственно отбирают наилучший из возможных методов. Отбор производится следующим образом. Для каждой пары «прогностический метод – стратегия его использования» на основе архивной выборки оцениваются значения К и R. Из множества пар отбираются пары, оптимальные по Парето, т. е. такие, у которых один из критериев экстремален на рассматриваемом множестве. Если подмножество оптимальных по Парето пар содержит только одну пару, то процедура отбора заканчивается, участие экспертов не требуется, а категорически вероятностный прогноз строится на основе этой пары. В противном случае эксперты отбирают из оптимального по Парето подмножества единственную пару. Если с точки зрения экспертов ни одна из указанных пар не является удовлетворительной, это подмножество отбрасывается, а на оставшемся подмножестве процедура повторяется вновь.

Достоинством метода прямых экспертных оценок является, с одной стороны, возможность существования решения, оптимального независимо от мнения экспертов, и, с другой стороны, если такого решения нет, простота процедуры экспертизы.

Рассмотрим применение всех трех описанных методов для построения конкретного категорически вероятного прогноза.

3.5.4 Категорически вероятностный прогноз посадки.

В качестве примера рассмотрим категорически вероятностный прогноз 12-часовой заблаговременности возможности посадки летательного аппарата. Прогноз разрабатывается на основе архивных данных АМОГ Воронеж на осенние месяцы 1971 – 1981 г.г. Во внимание принимались только внутримассовые процессы. Всего архив содержал 734 случая. Предполагалось, что возможность посадки определяется выполнением следующих условий: высота нижней границы облачности (НГО) H больше 200 м; дальность видимости D больше 2000 м; ско-

рость бокового ветра V меньше 12 м/с, а экономическая эффективность деятельности потребителя – средними затратами за фиксированный период времени (байесовский критерий эффективности).

В предварительный перечень предикторов были включены прогнозы НГО, видимости и скорости ветра различными методами. Для прогноза НГО использовались методы К. Г. Абрамович (метод 1) и И. В. Кошеленко (метод 2), скорости бокового ветра – гидродинамический (метод 1) и В. М. Ярковой (метод 2), дальности видимости Н. В. Петренко (метод 1) и И. В. Кошеленко (метод 2). При этом в силу ограниченности архива в качестве формулировок прогнозов брались формулировки «выше (ниже соответствующего минимума».

При построении метода категорически вероятностного прогноза необходимо решить две взаимосвязанные задачи: определить оптимальные прогностический метод и стратегию его использования. В рассматриваемом случае это означает определение оптимального сочетания методов (комплексация различных методов прогноза одной метеорологической величины не производилось) и набора формулировок, составленных этими методами прогнозов, при которых следует планировать полеты. Каждому сочетанию методов указанному набору формулировок соответствует своя матрица сопряженности n_{ij} (i = 1 по прогнозу принимается решение проводить полеты, i = 2 – не проводить, i = j – решение по метеоусловиям принято правильное, i ≠ j – неправильное). В терминах матрицы сопряженности критерии К и R имеют вид:

$$К = l_{11}n_{11} + l_{12}n_{12} + l_{22}n_{22}, \quad (3.31)$$

$$R = n_{12}, \quad (3.32)$$

Где $l_{iк}$ – соответствующие элементы матрицы затрат.

Для оценки значений К и R использовался архивный материал, приведенный в табл. 3.7. Каждый столбец таблицы соответствует различному набору формулировок прогностических методов. Каждая большая строка таблицы, соответствующая одному из восьми возможных сочетаний методов (например, 212 – соответствует второму методу прогноза дальности видимости и второму методу прогноза боковой скорости ветра), разбита на три строки: N_n – число данных сочетаний формулировок прогнозов рассматриваемыми методами, имеющихся в архиве; $N_Л$ - число случаев в архиве, в которых при данном наборе формулировок осуществились условия, допускающие возможность посадки; $Р_Л$ – оценка вероятности возможности посадки по метеоусловиям при данном наборе формулировок.

Категорически-вероятностный прогноз, минимизирующий К (без учета

безопасности полетов), строится следующим образом. На первом этапе по матрице затрат определяется P*. В перечень формулировок, допускающих посадку для каждого сочетания исходных методов, включаются формулировки, для которых $P_Л > P^*$. На втором этапе на этой основе для каждого сочетания методов строится матрица сопряженности и оценивается значение К. Методы, дающие наименьшее значение К, выбираются в качестве предикторов. При значениях предикторов. Соответствующих

$P_Л > P^*$, прогнозируется возможность посадки, в противном случае – невозможность. Пусть, например, P* = 0,6. Тогда список формулировок, допускающих возможность посадки для каждого сочетания исходных методов, имеет вид, приведенный в табл. 3.8. В третьей строке таблицы приведены значения:

$$\Delta = (1 - P^*)n_{12} - P^*n_{11} = \frac{К - l_{22} - Q(l_{21} - l_{22})}{l_{12} - l_{21} - l_{11} - l_{22}}, \qquad (3.33)$$

где Q – климатологическая вероятность осуществления условий, допускающих посадку.

Таблица 3.7

Статистика вероятности возможности посадки при различном сочетании методов прогноза H, V, и D

Сочетание методов	Величина	Формулировка исходных прогнозов							
		H$\bar{V}$D	$\bar{H}$VD	H$\bar{V}$D	HV$\bar{D}$	$\bar{H}\bar{V}$D	$\bar{H}$V$\bar{D}$	H$\bar{V}\bar{D}$	$\bar{H}\bar{V}\bar{D}$
H-1	$N_Л$	427	3	6	12	3	7	10	1
V-1	$N_П$	466	15	10	45	29	102	47	20
D-1	$P_Л$	0,91	0,20	0,60	0,26	0,10	0,07	0,21	0,05
H-2	$N_Л$	425	3	5	10	3	11	10	2
V-1	$N_П$	501	18	13	22	27	98	43	12
D-1	$P_Л$	0,84	0,17	0,38	0,45	0,11	0,11	0,23	0,16
H-1	$N_Л$	421	3	6	14	5	9	11	1
V-2	$N_П$	450	20	7	40	42	128	56	10
D-1	$P_Л$	0,94	0,15	0,86	0,35	0,17	0,07	0,19	0,10
H-1	$N_Л$	398	6	7	16	8	18	14	2
V-1	$N_П$	442	12	16	30	34	34	49	17
D-2	$P_Л$	0,90	0,50	0,44	0,53	0,24	0,13	0,29	0,11
H-2	$N_Л$	412	4	6	16	5	10	14	2
V-2	$N_П$	468	10	21	37	27	96	62	13
D-1	$P_Л$	0,88	0,40	0,28	0,43	0,19	0,10	0,22	0,15
H-1	$N_Л$	387	4	7	21	15	20	12	3
V-1	$N_П$	459	18	12	45	20	117	49	14
D-2	$P_Л$	0,84	0,22	0,58	0,46	0,75	0,17	0,24	0,21
H-1	$N_Л$	436	2	3	3	7	8	9	1
V-2	$N_П$	486	16	14	25	30	106	34	23
D-2	$P_Л$	0,89	0,13	0,21	0,12	0,23	0,08	0,26	0,04
H-2	$N_Л$	402	8	16	18	8	14	12	1
V-2	$N_П$	473	19	9	42	22	10	44	15
D-2	$P_Л$	0,84	0,42	0,66	0,43	0,36	0,13	0,27	0,06

Поскольку l_{ij} и Q не зависят от формулировок прогноза, минимум К соответствует минимуму Δ, а для выбора наилучшего сочетания и в качестве критерия можно пользоваться значением Δ. Из табл. 3.8 следует, что в качестве предикторов категорически вероятностного прогноза следует

выбрать прогнозы НГО по методу 1. При этом при формулировках HVD и HVD следует рекомендовать осуществлять полет, при других формулировках – не осуществлять.

Учет требований безопасности полетов может, как указывалось, изменить приведенные рекомендации. Рассмотрим указанные три метода учета этих требований.

Пример 5. Рассмотрим метод критерия полной эффективности.

Если $К_0$ существует (выполняется соотношение (3.30) для независимых оценок *c* и P*), описанная процедура повторяется для $К_0$. Пусть по оценкам экспертов P = 0,9. Данные, аналогичные данным табл.3.8, для такого значения пороговой вероятности приведены в табл. 3.9. из нее следует, что в качестве предикторов категорически вероятностного прогноза нужно выбрать прогноз НГО методом 1, скорости ветра методом 1 и дальности видимости методом 1, и прогнозировать возможность посадки только при формулировке HVD. Заметим, что если по оценкам экспертов $P_0^* = 0{,}95$, то при любых формулировках исходных прогнозов следует прогнозировать невозможность посадки.

Таблица 3.8

Выбор предикторов при категорически-вероятностном прогнозе для P* = 0,6

Сочетание методов	1.1.1	2.1.1	1.2.1	1.1.2	2.2.1	2.1.2	1.2.2	2.2.2
Набор формулировок	HVD HVD	HVD	HVD HVD	HVD	HVD	HVD HVD	HVD	HVD HVD
Δ	- 0.327	- 0.307	- 0.332	- 0.328	- 0.307	- 0.313	- 0.207	- 0.293

Таблица 3.9

Выбор предикторов при категорически-вероятностном прогнозе для P* = 0,9, по методу полной эффективности

Сочетание методов	1.1.1	2.1.1	1.2.1	1.1.2	2.2.1	2.1.2	1.2.2	2.2.2
Набор формулировок	HVD	-	HVD	HVD	-	-	-	-
Δ	0.518	0	-0.512	-0.458	0	0	0	0

Рассмотрим метод допустимых значений R.

При таком подходе в набор формулировок, допускающих возможность посадки для каждого сочетания методов, включаются только такие, для которых по архивным оценкам n_{12} меньше R_m, определенного экспертами. Затем производится расчет (P* определяется по формуле 3.28 из матрицы экономических затрат) и по минимуму определяется наилучшее сочетание исходных методов. Так, если R_m = 0,05 а P* = 0,6, как в первом случае, по данным табл. 3.10 определяем, что в качестве предикторов категорически вероятностного прогноза следует использовать исходные методы 1,2 и 1 соответственно и предсказывать возможность посадки при формулировках HVD и H$\bar{V}$D.

Таблица 3.10

Выбор предикторов при категорически-вероятностном прогнозе для P* = 0,6 по методу допустимых значений

Сочетание методов	1.1.1	2.1.1	1.2.1	1.1.2	2.2.1	2.1.2	1.2.2	2.2.2
Набор формулировок	HVD	HVD	HVD HVD	HVD	HVD	HVD	HVD	HVD
R	0.053	0.104	0.041	0.060	0.076	0.098	0.191	0.097
Δ	-0.282	-0.307	-0.328	-0.300	-0.307	-0.277	-0.207	-0.317

Рассмотрим метод прямых экспертных оценок.

В этом случае построение метода категорически вероятностного прогноза заключается в определении оптимальных по Парето методов с последующей их, в случае необходимости, экспертной оценкой.

При фиксированном значении P* минимальное значение Δ обеспечивается включением для каждого сочетания исходных методов в набор формулировок, допускающих посадку, таких, для которых $P_Л \geq P$. Минимальное значение R достигается. Очевидно, при отсутствии полетов вообще. Если подобное решение априори неприемлемо, оно изначально отбрасывается и не включается в анализируемое множество, а для каждого набора исходных методов определяется сочетание формулировок, обеспечивающее минимальное для данного сочетания значение R. В обоих случаях рассчитывается значение Δ и R. Соответствующие результаты приведены в табл. 3.11. На основании такой таблицы определяется два метода категорически-вероятностного прогноза, оптимальных по Парето (один из них обеспечивает минимум R, другой - Δ). Если оба метода совпадают – задача решена, если нет – выбор предоставляется экспертам. В рассматриваемом случае обоим оптимальным по Парето методам соответствует выбор в качестве предикторов формулировок прогнозов ис-

ходными методами 1,2 и 1. Выбор же набора формулировок, при которых прогнозируется возможность посадки (HVD или HVD и H$\bar{V}$D), должен быть произведен экспертами или лицом, принимающим решение на основе данных табл. 3.11. Существенно, что, поскольку любые другие методы категорически-вероятностного прогноза обеспечивают худшие значения обоих критериев, выбор экспертов ограничен только этими двумя методами.

Таблица 3.11

Сочетание методов	1.1.1	2.1.1	1.2.1	1.1.2	2.2.1	2.1.2	1.2.2	2.2.2
Набор формули-Ровок для min Δ	HVD H$\bar{V}$D	HVD	HVD H$\bar{V}$D	HVD	HVD	HVD H$\bar{V}$D	HVD	HVD H$\bar{V}$D
P/Δ	$\frac{0.063}{-0.327}$	$\frac{0.104}{-0.307}$	$\frac{0.041}{-0.328}$	$\frac{0.060}{-0.328}$	$\frac{0.076}{-0.307}$	$\frac{0.105}{-0.313}$	$\frac{0.191}{-0.207}$	$\frac{0.101}{-0.293}$
Набор формули-Ровок для min Δ	HVD	HVD	HVD	HVD	HVD	HVD	HVD	HVD
P/Δ	$\frac{0.053}{-0.328}$	$\frac{0.104}{-0.307}$	$\frac{0.040}{-0.328}$	$\frac{0.060}{-0.300}$	$\frac{0.076}{-0.307}$	$\frac{0.098}{-0.277}$	$\frac{0.191}{-0.207}$	$\frac{0.097}{-0.317}$

3.6 Выбор стратегии при поэтапном планировании.

В практической деятельности часть планируемых потребителем метеоинформации задач выполняется с учетом результатов достигнутых на предыдущих этапах. При этом платежная матрица видоизменяется от этапа к этапу, что накладывает определенные сложности на использование алгоритмов, рассмотренных выше. В связи с тем, что из-за априорной неопределенности платежной матрицы при поэтапном планировании затруднено использование классических методов теории статистических решений, а вероятностный характер метеорологических прогнозов исключает возможность применения в этих целях алгоритмов планирования при детерминированных условиях, характерных, например, для теории расписаний, то в настоящем параграфе будут рассмотрены алгоритмы метеорологического обеспечения поэтапного выполнения планируемых потребителем задач при заданных ограничениях на их выполнение.

В общем виде проблема оптимального использования метеорологической информации при поэтапном планировании может быть сформулирована следующим образом.

Потребителю необходимо реализовать программу, предусматривающую выполнение некоторых народнохозяйственных или военных задач $\langle B_1, B_2, \ldots, B_q \rangle = \bar{B}_{<q>}$. Планирование потребителем его деятельности, направленной

на реализацию программ, производится поэтапно: каждому этапу ставится в соответствие элементарный календарный период (ЭКП), на который в конце предыдущего ЭКП разрабатывается очередной и «оперативный» план функционирования потребителя.

При разработке планов считаются известными:

– ЭКП в течение которых должна выполняться каждая программная задача $\langle n_1,\ n_2, \dots, n_q \rangle = \bar{n}_{<q>}$;

– не метеорологические ограничения на функционирование (требуемая последовательность выполнения задач, допустимость их совместного выполнения в одном ЭКП и выполнения незапланированных задач и т. п.);

– метеорологические ограничения, определяющие метеорологические условия, при которых возможно выполнение каждой программной задачи - $\langle \mathrm{M}_1,\ \mathrm{M}_2, \dots, \mathrm{M}_q \rangle = \bar{\mathrm{M}}_{<q>}$.

Потребителю регулярно представляются (в категорической или вероятностной форме) детализированные по ЭКП прогнозы метеорологических условий «выполнимости» программных задач $\bar{\mathrm{M}}_{<q>}$. Следует подчеркнуть, что для эффективного планирования потребителю необходимы прогнозы именно метеорологических условий выполняемости программных задач, а не отдельных метеорологических величин или погодных явлений. Разработка таких прогнозов может производится как с использованием статистических связей $\mathrm{M}_{<q>}$ с исходными характеристиками состояния атмосферы, так и на основе комплексации формулировок «стандартных» прогнозов.

Требуется разработать рекомендации по оптимальному (т. е. доставляющему экстремум заданному показателю эффективности) использованию прогнозов различных видов в процессе поэтапного планирования. Показатель эффективности оперативного планирования определяется исходя из требований программы. Им может быть вероятность выполнения программных задач к установленному сроку или математическое ожидание времени, затраченного на выполнение таких задач. Рассмотрим более подробно методы, основанные на поиске экстремумов этих показателей эффективности.

3.6.1 Метод нахождения минимума математического ожидания времени, затраченного на выполнение задачи

При выполнении определенного класса задач потребитель метеорологический информации сталкивается с задачей реализации программы своей деятельности в минимально возможные сроки.

Рассмотрим задачу осуществления программы авиационных перевозок в различные пункты назначения.

Пусть стратегия поэтапного планирования реализации программы предусматривает выполнение потребителем задач двух типов: B_1 и B_2. Принимается, что планы разрабатываются на ближайший ЭКП, причем все ЭКП имеют одинаковую продолжительность, количество ЭКП, на протяжении которых необходимо выполнять задачу B_1, равно n_1, а задачу B_2 - n_2. В течение одного ЭКП может решаться задача только одного типа и только предусмотренная планом. На каждый ЭКП разрешено планирование любой задачи, независимо от того, какие задачи решились на предыдущих ЭКП.

Выполнение в данном ЭКП задачи B_i, допустимо только при осуществлении в этот период метеорологических условий M_i. Вероятность их осуществления считается для всех ЭКП одинаковой и независящей от ранее наблюдавшихся условий.

Эффективность планирования оценивается величиной математического числа ЭКП $\bar{N}$, необходимых для реализации программы.

Под прогнозом понимается предсказание сочетаний метеорологических условий выполнимости задач M_1M_2, $M_1\bar{M}_2$, $\bar{M}_1M_2$, $\bar{M}_1\bar{M}_2$.

Общая формула для расчета $\bar{N}$ при реализации простейших программ $< n_1, 0 >$ и $< 0, n_2 >$ при стратегии планирования, когда на каждый ЭКП при ревой программе планируется выполнение задачи B_1, а при второй - B_2, имеет вид:

$$\bar{N}_{<n,0>} = \sum_{s=0}^{\infty} C_{n_1+s-1}^{2}(n_1 + s) + (P_{<M_1\bar{M}_2>} + P_{<M_1M_2>})^{P_1} \text{ X}$$
$$\text{X } (P_{<M_1M_2>} + P_{<\bar{M}_1M_2>})^{s} = \frac{n_2}{P_{<M_1\bar{M}_2>} + P_{<M_1M_2>}} \qquad (3.34)$$

и аналогично

$$\bar{N}_{<0,n_2>} = \frac{n_2}{P_{<\bar{M}_1M_2>} + P_{<M_1M_2>}}, \qquad (3.35)$$

где $P_{<>}$ – вероятность осуществления сочетания метеорологических условий, указанных в нижнем индексе;

C_k^1 – число сочетаний из k элементов по 1.

Пример 6. При планировании используются «оперативные» (не идеальные) прогнозы на ближайший ЭКП. Известна характеризующая оправдываемость этих прогнозов, матрица сопряженности (табл. 3.12) и климатические вероятности.

Таблица 3.12

Матрица сопряженности для оперативных прогнозов

Предсказанные условия	Осуществившиеся условия	
	M_1	M_2
$M_1\bar{M}_2$	P_{11}	P_{12}
$\bar{M}_1M_2$	P_{21}	P_{22}
M_1M_2	P_{31}	P_{32}
$\bar{M}_1\bar{M}_2$	P_{41}	P_{42}

В табл. 3.12 P_{ij} – вероятность осуществления условий M_j, при i-той формулировке прогноза.

Предполагается, что прогнозы являются «несмещенными» (вероятность каждой формулировки равна климатической вероятности соответствующих метеорологических условий) и не «зеркальными» (при прогнозах $M_1\bar{M}_2$ и M_1M_2 целесообразно планировать выполнение задач соответственно B_1 и B_2, а не B_2 и B_1). Для выбора оптимальной стратегии использования таких прогнозов можно ограничиться рассматриванием стратегий, представленных в табл. 3.13.

Математическое ожидание числа ЭКП, требуемого для реализации программы при указанных стратегиях, рассчитывается по формуле:

$$\bar{N}_{<n_1,n_2>}(S_i) = \frac{1}{a_i + b_i}\left(a_i \bar{N}_{<n_1 - i\, n_2>} + b_i \bar{N}_{<n_1,n_2-1>} + 1\right) \qquad (3.36)$$

где i = 1 (1) 4.

Таблица 3.13

Стратегия планирования с использованием оперативных прогнозов

Предсказанные условия	Планируемые задачи при стратегиях			
	s_1	s_2	s_3	s_4
$M_1\bar{M}_2$	B_1	B_1	B_{12}	B_1
$\bar{M}_1M_2$	B_2	B_2	B_2	B_2
M_1M_2	B_1	B_2	B_1	B_2
$\bar{M}_1\bar{M}_2$	B_1	B_1	B_2	B_2

Значение параметров a и b даны в табл. 3.14.

Таблица 3.14

Значения коэффициентов a и b в формуле (3.36) для различных стратегий

S	a	b
S_1	$P_{<M_1\bar{M}_2>} * P_{11} + P_{<M_1M_2>} * P_{31} + P_{<M_1M_2>} * P_{41}$	$P_{<\bar{M}_1M_2>} * P_{22}$
S_2	$P_{<M_1\bar{M}_2>} * P_{11} + P_{<\bar{M}_1\bar{M}_2>} * P_{41}$	$P_{<\bar{M}_1M_2>} * P_{22} + P_{<M_1M_2>} * P_{33}$
S_3	$P_{<M_1\bar{M}_2>} * P_{11} + P_{<M_1M_2>} * P_{31}$	$P_{<\bar{M}_1M_2>} * P_{22} + P_{<\bar{M}_1\bar{M}_2>} * P_{42}$
S_4	$P_{<M_1\bar{M}_2>} * P_{11}$	$P_{<\bar{M}_1M_2>} * P_{22} + P_{<M_1M_2>} * P_{32} + P_{<\bar{M}_1\bar{M}_2>} * P_{42}$

Расчеты для формулы (3.36) производятся последовательно для программ $<1,1><2,1>$, …, $<n,n>$, и для каждой программы определяется стратегия обеспечивающая минимум $\bar{N}_{<>}$.

Данная оптимизационная задача решается методом динамического программирования.

3.6.2 Метод нахождения максимума вероятности выполняемой задачи по метеорологическим условиям

В отличие от рассматриваемого ранее примера существует класс потребителей метеорологической информации, которые желают за определенный срок реализовать программу своей деятельности с максимальной вероятностью.

Рассмотрим задачу проведения исследований в атмосфере с помощью различного типа автоматических аэростатов. Причем, при выполнении исследовательской задачи автоматический аэростат любого типа может использоваться только один раз.

Пусть стратегия поэтапного планирования реализации программы предусматривает выполнение потребителем t-типов задач ($B_1,\ B_2, \dots, B_t$). За календарный период, состоящий из № ЭКП (этапов) требуется при имеющихся материальных ресурсах К($К = \sum_{i=1}^{t} K_i$) с максимальной вероятностью max P(N, K, B), выполнить В – задач ($В = \sum_{i=1}^{t} B_i$). Выполнение в данном ЭКП задачи B_i, допустимо только при осуществлении в этот период метеорологических условий M_i.

Под прогнозом понимаются статистические характеристики наборов формулировок осуществления предсказанных «традиционными» способами метеорологических М условий задачи.

Таблица 3.15

Таблица сопряженности между предсказанными и осуществившимися метеорологическими условиями

Осуществившиеся условия	Предсказанные условия						
	F_1	F_2	…	F_k	…	F_s	Σ
$Э_1$	n_{11}	n_{12}	…	n_{1k}	…	n_{1s}	n_{10}
$Э_2$	n_{21}	n_{22}	…	n_{2k}	…	n_{2s}	n_{20}
…	…	…	…	…	…	…	…
$Э_i$	n_{i1}	n_{i2}	…	n_{ik}	…	n_{is}	n_{i0}
…	…	…	…	…	…	…	…
$Э_t$	n_{t1}	n_{t2}	…	n_{tk}	…	n_{ts}	n_{t0}
Σ	n_{01}	n_{02}	…	n_{0k}	…	n_{0s}	n_{00}

Пусть

q_i - вероятность аварии при выполнении B_i - задачи в неблагоприятных метеорологических условиях;

p_{ij} - повторяемость осуществления прогнозируемых метеоусловий для выполнения B_i – задачи при j-том наборе формулировок прогноза;

c_{ij} - повторяемость j-того набора формулировок прогноза метеоусловий, благоприятных для выполнения B_i – задачи.

Наборы формулировок прогноза для выполнения i-того вида задачи определяются по таблице сопряженности (табл. 3.15) следующим образом:

Определяются p_{ij} и c_{ij}, рассчитанные для одной из формул S формулировок прогноза. Например,

$$p_{i1} = \frac{n_{i1}}{n_{01}}; \quad c_{i1} = \frac{n_{0j}}{n_{00}};$$
$$p_{i2} = \frac{n_{i2}}{n_{02}}; \quad c_{i2} = \frac{n_{02}}{n_{00}};$$
$$\dots \qquad \dots$$
$$p_{is} = \frac{n_{is}}{n_{0s}}; \quad c_{is} = \frac{n_{0s}}{n_{00}};$$

при этом общее количество наборов формулировок составляет

$$c_s^1 = \frac{s!}{1!(s-1)!} = s; \tag{3.37}$$

определяются p_{ij} и c_{ij}, рассчитанные для двух из S формулирвок прогноза. Например,

$$p_{i(s+1)} = \frac{n_{i1} + n_{i2}}{n_{01} + n_{02}}; \qquad s_{s(s+1)} = \frac{n_{01+n02}}{n_{01}};$$
$$p_{i(s+2)} = \frac{n_{i2} + n_{i3}}{n_{02} + n_{03}}; \qquad c_{i(s+2)} = \frac{n_{02} + n_{03}}{n_{00}};$$
$$\dots \qquad \dots \qquad \dots$$
$$p_{i(s+\lambda)} = \frac{n_{i1} + n_{is}}{n_{01} + n_{0s}}; \qquad s_{s(s+\lambda)} = \frac{n_{01} + n_{0s}}{n_{00}};$$

при этом количество наборов формулировок составляет

$$c_s^2 = \frac{s!}{2!(s-2)!}; \tag{3.38}$$

Затем определяется p_{ij} и c_{ij}, рассчитанные для сочетаний трех, четырех, …, из S формулировок прогноза, общее количество которых составляет:

$$H = \sum_{k=1}^{s} \frac{s!}{k!(s-k)!}. \tag{3.39}$$

Очевидно, что при выполнении программы

$$\left.\begin{array}{ll} P(N, K, O) = 1 & \forall N, K \\ P(N, K, B) = 0 & \forall N < B \\ P(N, R, B) = 0 & \forall K_i < B_i \end{array}\right\} \tag{3.40}$$

При принятии решения на выполнение задачи на каждом из N-этапов возможны следующие исходы;

- с вероятностью p_{1j} * c_{1j} ожидается выполнение задачи 1-го типа (j ∈ h), j-тый набор включает n-формулировок прогноза (n ∈ s); с вероятностью $c_{i(j+1)}$ * $p_{i(j+i)}$ ожидается выполнение задачи 2-го типа, причем (j +1)-й набор может включать только (s – n) – формулировок прогноза, и т. д. для всех

i = $\overline{1,t}$, j = $\overline{1,h}$;

- с вероятностью q_{ij} * $c_{ij}(1 - p_{1j})$ ожидается невыполнение задачи в неблагоприятных метеоусловиях с потерей K_i-го типа материальных средств, с помощью которых выполняется задача (для всех i = $\overline{1,t}$, j = $\overline{1,h}$);

- с вероятностью $1 - \sum_{j-t} c_j$ с ожидается решение потребителя не выполнять задачи.

Общее рекуррентное соотношение для решения оптимизационной задачи имеет вид:

$$\begin{aligned} &P(N, K, B) = \sum_{i=1}^{t}\sum_{j=1}^{h} c_j\,[p_{ij}P(N-1,\ K_i-1, B_i-1) + \\ &+ q_i(1 - p_{ij}) * P(N-1, K_i - 1, B) + \\ &+ (1- q_i)(1 - p_{ij})\,P(N-1, K_i - 1, B_i - 1)] + \\ &+ (1 - \sum_{j=1}^{h} c_j)\,P(n-1, K, B). \end{aligned} \qquad (3.41)$$

Данная оптимизационная задача решается методом динамического программирования.

Пример 7. Пусть календарный период, состоящий из N (этапов) потребителю со стратегией полного доверия прогнозам метеоролога, требуется с максимальной вероятностью max P(N, K, B) выполнить программу, состоящую из выполнения M-задач одного типа (t = 1), для чего выделяется K материальных средств. Пусть метеоролог имеет в наличии статистические характеристики p_j и c_j (j = $\overline{1,L}$) L-способов прогноза.

Общее рекуррентное соотношение (3.41) при решении задачи примет вид:

$$\begin{aligned} &P(N, K, B) = \max_j c_j[p_j P(N-1, K-1, B-1) + \\ &+ q(1 - p_j)\,P\,(N-1, K-1, B) + \\ &+ (1 - q)(1 - p_j)\,P\,(N-1, K-1, B-1) + \\ &+ (1 - c_j)\,P\,(N-1, K, B)]. \end{aligned} \qquad (3.42)$$

Решение оптимизационной задачи достигается методом динамического программирования. При этом первый ЭКП (на первом этапе) n = 1, K = 1,

B = 1 с учетом соотношений (3.40) выражение (3.42) имеет вид:

$$P_j(1.1.1) = c_j p_j + c_j(1 - p_j)*(1 - q) = c_j(1 - q(1 - p_j)). \qquad (3.43)$$

В этом случае оптимальным следует считать способ прогноза, при котором достигается

$$\max_j c\,(1 - q(1 - p)).$$

Дальнейшие расчеты по формуле (3.42) производятся последовательно для программ (2.1.1), (3.1.1), …, (N,1,1), …, (2.2.1), (3.2.1), …, (N,K,1), …, (2.2.2), (3.2.2), …, (N,K,B) и для каждой программы определяется способ прогноза, обеспечивающий максимум P(N,K,B).

4 СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

4.1 постановка задачи проверки гипотез

Гипотезой в общем случае называется предположение о некоторых свойствах изучаемых явлений. При статистической обработке метеорологической информации в качестве таких предположений можно рассматривать следующие: о виде закона распределения случайной величины, о параметрах закона распределения, о случайности получаемого результат и т. п. Всякие предположения, выдвинутые на основании анализа результатов наблюдения, называются статистическими гипотезами.

Выдвинутая гипотеза в результате проверки может подтвердиться или не подтвердиться. Поэтому наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу, которая может быть принята в том случае, если первая не подтвердится. Для отличия этих гипотез друг от друга применяется специальная терминология и вносятся соответствующие обозначения. Выдвинутую гипотезу принято называть нулевой или основной и обозначать символом H_0, а гипотезу, противоречащую нулевой, - конкурирующей или альтернативной и обозначать символом H_1.

Пусть, например, выдвинуто предположение о том, что исследуемая характеристика подчиняется равномерному закону распределения. Тогда это предположение является нулевой гипотезой. Альтернативных к ней гипотез может быть выдвинуто несколько. Одна из них может состоять в том, что эта характеристика не подчиняется равномерному закону распределения.

Для удобства записи используются обозначения, принятые в математике. Пусть нулевая гипотеза состоит в предположении, что математические ожидания двух нормально распределенных совокупностей

$\hat{x}$ и $\hat{y}$ равны, а альтернативная гипотеза состоит в том, что они не равны. Запись этих гипотез осуществляется следующим образом: H_0: $m_x = m_y$;

H_1: $m_x \neq m_y$.

Статистические гипотезы принято подразделять на простые и сложные.

Простой называют гипотезу, содержащую только одно предположение, например, предположение о том, что завтра температура днем будет 15° (H_0: t = 15 °).

Сложной называют гипотезу, содержащую конечное или бесконечное число предположений. Например, гипотеза H_0: $t > 15°$ состоит из бесчисленного множества простых гипотез вида H_i: $t = b_i$, где b_i – любое число, превосходящее 15.

Любая статистическая гипотеза, как предположение, должна подвергаться проверке (испытанию). Задачи проверки гипотез можно разделить на несколько классов, отличающихся друг от друга как по форме, так и по методам решения. Прежде всего, все задачи проверки гипотез делятся на параметрические, когда параметры закона распределения известны, и непараметрические, когда параметры неизвестны.

В свою очередь, каждый из данных классов задач содержит следующие подклассы.

Задачи согласия. Данные задачи включают в себя задачи соответствия (согласия) вида закона распределения или значений параметров распределения, выдвинутых в качестве предполагаемых, с законом распределения или параметрами закона распределения реальной случайной величины. Формулировка данных задач имеет следующий вид. Нулевая гипотеза: H_0: $Q_{\hat{x}} = S_{\hat{x}}$. конкурирующие гипотезы: H_1^1: $Q_{\hat{x}} < S_{\hat{x}}$; H_1^2: $Q_{\hat{x}} > S_{\hat{x}}$;

H_1^3: $Q_{\hat{x}} \neq S_x$. Здесь $Q_{\hat{x}}$ и $S_{\hat{x}}$ обозначают символ закона распределения, то $Q_{\hat{x}}$ = = $F_{\hat{x}}$, $S_{\hat{x}} = F_{\hat{x}}$, где $F_{\hat{x}}$ - истинный закон распределения исследуемой случайной величины, $F_{\hat{x}}$ – гипотетический закон распределения. Для нашего случая получим: H_0: $F_{\hat{x}} = F_{\hat{x}_r}$; H_1: $F_{\hat{x}} \neq F_{\hat{x}_r}$.

При проверке согласия параметров распределения альтернативные гипотезы могут выдвигаться в форме H_1^1 и H_2^2, т. е. в форме гипотез, содержащих бесчисленное множество предположений.

Задачи независимости. Эти задачи возникают в тех случаях, когда необходимо проверить, являются ли компоненты некоторого случайного вектора $\hat{X}_{<n>}$ независимыми. Из теории вероятностей известно, что вероятность произведения независимых случайных величин равна произведению вероятностей этих величин. Поэтому, если компоненты вектора $\hat{X}_{<n>} = \langle \hat{X}_1, \hat{X}_{2,} \dots, \hat{X}_n \rangle$ независимы, то функция распределения вектора будет равна произведению функций распределения компонент.

$$F_{\hat{x}_{<n>}}(X_{<n>}) = \prod_{i=1}^{n} F_{\hat{x}_i}(x_i).$$

Если же это равенство не выполняется, то компоненты являются зависимыми величинами. Поэтому задачу проверки гипотез о независимости можно сформулировать следующим образом:

$$H_0: F_{\hat{x}_{<n>}}(X_{<n>}) = \prod_{i=1}^{n} F_{\hat{x}}(X_i),$$

$$H_1: F_{\hat{x}_{<n>}}(X_{<n>}) \neq \prod_{i=1}^{n} F_{\hat{x}_i}(x_i).$$

Задачи проверки выборки. Данные задачи возникают при необходимости проверки того факта, что полученная выборка является простой, т. е. различные

части выборки подчинены одному и тому же закону распределения. Задачи такого типа формулируются в виде соотношений:

$$H_0: F_{\hat{x}_{<n>}}(x_1,\ x_{2,}\ \dots, x_n) = F_{\hat{x}}(x_i),$$

$$H_1: F_{\hat{x}_{<n>}}(x_1,\ x_{2,}\ \dots, x_n) \neq F_{\hat{x}}(x_i).$$

Из постановки задач проверки гипотез следует, что проверку гипотез можно рассматривать как частный случай теории статистических решений и, следовательно, для ее решения полностью применимы методы этой теории. Наиболее часто методы проверки гипотез используются для решения задачи согласия. Общий подход к решению задачи проверки гипотез включает в себя 5 основных этапов:

На первом этапе решения задачи определяется множество всех возможных решений (пространство решений). В проверке гипотез это множество обычно состоит из двух элементов: решения d_1, состоящего в принятии гипотезы H_0, и решении d_2, состоящего в отклонении гипотезы H_0, или принятии гипотезы H_1.

На втором этапе должно быть условлено решающее правило, по которому определяются условия принятия или не принятия нулевой гипотезы. Так, как речь идет об аналитическом решении задачи, то эти условия должны заключаться в выборе некоторой величины, представляющей собой функцию элементов выборки, связанной с нулевой и альтернативной гипотезами и зависящей от условий проведения эксперимента. Эту величину принято называть показателем согласованности гипотезы и обозначать символом $\widehat{U}$.

На третьем этапе выбирается ряд правил, по которым устанавливается, при каких значениях показателя согласованности гипотеза принимается, а при каких отвергается, т. е. выбирается критерий проверки гипотез. Иногда этот критерий называют критерием согласия или критерием соответствия.

На четвертом этапе множество возможных значений величины показателя согласованности в соответствии с принятым критерием разбивается на два множества таким образом, что попадание возможного значения показателя согласованности в одно из этих подмножеств означает принятие гипотезы H_0, а в другое - отклонение гипотезы H_0.

На пятом этапе проводятся испытания, вычисляется величина U и определяется, к какому из подмножеств относится эта величина. На основании этого принимается решение о приеме или отклонении гипотезы H_0.

Из описаний процедуры следует, что в решении задачи проверки гипотез важнейшим элементом является выбор показателя согласованности и критерия проверки гипотез, которые должны обладать определенными свойствами.

4.2 Показатель согласованности и его свойства

Как уже указанно в подразд. 4.1, основными задачами проверки гипотез являются задачи согласия, в которых проверяются предположения о соответствии вида закона распределения или значений параметров распределения, полученных по экспериментальным данным, некоторому теоретическому закону с соответствующими параметрами. Если удается установить, что реальный закон близок к некоторому теоретическому, то его легко описать аналитически. В общем случае вывод аналитического описания плотности или функции распределения представляет собой довольно трудную задачу, которую не всегда удается решить. Поэтому выгоднее найти известный теоретический закон, к которому близко экспериментальный. По этой причине показатель согласованности должен являться функцией, как данных результатов наблюдения, так и некоторых гипотетических данных.

Показателем согласованности или статистической характеристикой гипотезы называется случайная величина U, являющаяся функцией гипотетических данных и данных результатов наблюдений и представленная для проверки нулевой гипотезы. Конкретный вид показателя согласованности для различных гипотез может быть различным. Так при проверке гипотезы о законе распределения он может задаваться следующими способами:

- в виде зависимости от гипотетической функции распределения $F_{\hat{x}_r}(x)$, т. е. функции распределения закона, выдвинутого в качестве нулевой гипотезы, и статистической функции распределения $F^*_{\hat{x}}(x)$, полученной экспериментально:

$$U = f_1(F_{\hat{x}_r}(x), F^*_{\hat{x}}(x)) \quad (4.1)$$

- в виде зависимости от гипотетической вероятности P и относительной частоты P*, полученной в результате проведения эксперимента:

$$U = f_1(P, P^*). \quad (4.2)$$

При проверке гипотезы о равенстве математических ожиданий двух независимых случайных величин $\hat{x}$ и $\hat{y}$ показатель согласованности может выбираться в виде различных зависимостей от начальных и центральных моментов первого, и второго и более высоких порядков случайных величин $\hat{x}$ $\hat{y}$:

$$U = f_3(\hat{V}^*_1[\hat{x}], \hat{V}^*_1[\hat{y}], \hat{\mu}^*_2[\hat{x}], \hat{\mu}^*_2[\hat{y}]) \quad (4.3)$$

Однако, несмотря на такое разнообразие, в любом случае показатель согласованности должен удовлетворять ряду требований. Так как показатель согласованности – величина случайная, то эти требования формулируются применительно к его закону распределения и состоят в следующем:

Закон распределения показателя согласованности должен зависеть от нулевой и альтернативной гипотез, а также от условий проведения эксперимента.

Так, в формуле (4.1) эта зависимость представлена наличием как гипотетической, так и статистической функцией распределения в качестве аргументов f_1.

Показатель согласованности должен представлять собой случайную величину, точное или приближенное распределение которой известно. В настоящее время наиболее распространен выбор показателей согласованности, распределенных по нормальному закону, законам хи-квадрат. Стьюдента, Фишера-Снедекора. В литературе по математической статистике распространено обозначение показателей согласованности, имеющих различные законы распределения, разными символами. Так, показатели согласованности, распределенные по нормальному закону, обозначают через U или Z, по закону хи-квадрат – через c, по закону Стьюдента – через χ^2, по закону Фишера – через F или v^2.

Закон распределения показателя согласованности должен быть инвариантен к виду закона распределения исследуемой случайной величины, т. е. не должен изменяться при смене закона распределения исследуемой величины. Именно данное обстоятельство определило широкое распространение показателей согласованности, имеющих указанные выше законы распределения.

Закон распределения показателей согласованности должен определяться при использовании минимума априорных сведений, так как возможность получения достоверных сведений до опыта существенно ограничена.

Закон распределения показателя согласованности $\widehat{U}$ должен быть критичен по отношению к проверяемой гипотезе. Данное требование означает, что условные плотности распределения $\varphi_{\widehat{U}/H_0}(\mathrm{U})$ и $\varphi_{\widehat{U}/H_1}(\mathrm{U})$ должны существенно отличаться друг от друга.

На рис. 4.1 изображены кривые условных плотностей распределения двух различных показателей согласованности U_1 и U_2 при нулевой (сплошные кривые) и альтернативной (пунктирные кривые) гипотезах. Из сравнения кривых видно, что применение показателя согласованности U_1 предпочтительнее, так как он позволяет более уверенно различать гипотезы H_0 и H_1, чем показатель U_2. Действительно, при одном и том же значении $\mathrm{U} = U_1^* = U_2^*$, вероятность отнесения его к нулевой гипотезе для U, значительно выше, чем для U_2.

Для проверки гипотезы по данным экспериментальной выборки вычисляют частные значения входящих в показатель согласованности величин и, таким образом, получают частное значение показателя согласованности гипотезы. Это значение, вычисленное по данным выборки, называют наблюдаемым значением показателя согласованности U*.

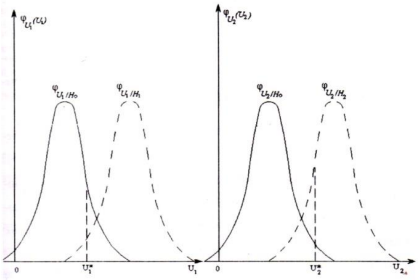


Рис. 4.1 Кривые условных плотностей распределения двух различных показателей согласованности U_1 и U_2.

4.3 Методы задания критической области

В соответствии с требованиями теории статистических решений при проверке гипотез необходимо задание решающего правила, т. е. метода разбиения множества U возможных значений показателя согласованности $\widehat{U}$ на два подмножества: U_0, при попадании в которое наблюдаемого значения U* показателя $\widehat{U}$ нулевая гипотеза принимается, и подмножество U_1, при попадании в которое наблюдаемого значения U* показателя $\widehat{U}$ нулевая гипотеза отвергается. Область, соответствующую подмножеству U_0, будем называть областью допустимых значений (областью принятия гипотезы H_0) и обозначать символом D, а область, соответствующую подмножеству U_1 - критической областью показателя согласованности $\widehat{U}$ и обозначить символом Q.

Поскольку показатель согласованности $\widehat{U}$ - одномерная случайная величина, то все ее возможные значения принадлежат некоторому интервалу. Поэтому область допустимых значений D и критическая область Q также являются интервалами и, следовательно, существуют точки, которые их разделяют. Эти точки называют критическими точками (границами). Таким образом, задание критической области сводится к заданию критических точек.

Сущность задания критических точек сводится к следующему.

Рассмотрим случайные события (для упрощения записи в данном параграфе «крышечки» над буквами ставить не будем):

A – верна гипотеза H_0;

$\overline{A}$ – верна гипотеза H_1;

B – наблюдаемое значение U* попало в область D;

$\overline{B}$ – наблюдаемое значение U* попало в область Q.

Тогда при принятии решения возможен один из следующих исходов:

A∩B – верна гипотеза H_0 и принято решение о ее справедливости;

$A \cap \overline{B}$ – верна гипотеза H_0, а принято решение о справедливости гипотезы H_1;

$\overline{A} \cap B$ – верна гипотеза H_1, а принято решение о справедливости гипотезы H_0;

$\overline{A} \cap \overline{B}$ – верна гипотеза H_1, и принято решение о ее справедливости.

Легко видеть, что исходы $A \cap \overline{B}$ и $\overline{A} \cap B$ являются ошибочными решениями. Исход $A \cap \overline{B}$ принято называть ошибкой первого рода, а исход $\overline{A} \cap B$ – ошибкой второго рода. Таким образом, под ошибкой первого рода понимается принятие решения об отклонении нулевой гипотезы в случае, если она является правильной, а под ошибкой второго рода – решение о принятии нулевой гипотезы в случае, если в действительности она не верна.

Пример 1. Пусть синоптик в качестве нулевой гипотезы выдвигает предположение о том, что днем будет гроза. Командир на основе этого прогноза может принять решение на проведение или «отбой» полетов.

Если он поверит прогнозу «отобьет» полеты, и гроза на самом деле осуществится, то решение будет принято правильно. Если при этом. В силу неидеальности прогнозов, грозы не будет, то будет допущена ошибка второго рода (принято решение об осуществлении гипотезы H_0, а верна гипотеза H_1). Если синоптик прогнозирует отсутствие грозы, проводятся полеты, а гроза возникает, то совершается ошибка первого рода. Так как рассмотренные события являются случайными, то им могут быть поставлены в соответствие вероятности наступления данных событий, а именно:

P_{11} - вероятность наступления $A \cap B$;

P_{12} - вероятность наступления $A \cap \overline{B}$;

P_{21} - вероятность наступления $\overline{A} \cap B$;

P_{22} - вероятность наступления $\overline{A} \cap \overline{B}$.

Значения P_{11}, P_{12}, P_{21}, P_{22} могут быть вычислены как вероятности попадания случайной величины $\hat{U}$ в области D и Q следующим образом.

Пусть законы распределения показателя согласованности $\hat{U}$ заданы при условиях, что справедлива нулевая или альтернативная гипотеза, в форме плотности распределения $\varphi_{\hat{U}/H_0}(U)$ и $\varphi_{\hat{U}/H_1}(U)$, а границей данных областей является точка U. Предположим, что взаимное расположение кривых распределений $\varphi_{\hat{U}/H_0}(U)$ и $\varphi_{\hat{U}/H_1}(U)$, а также областей D и Q имеет вид, представленный на рис. 4.2.

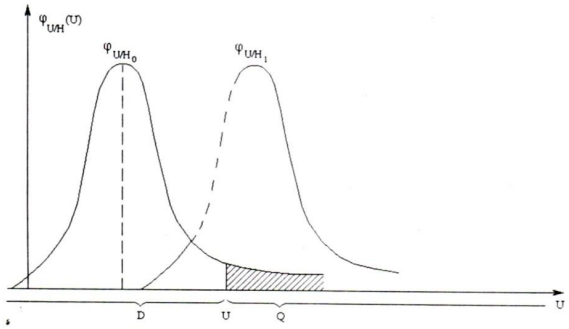


Рис. 4.2

Тогда вероятности P_{11}, P_{12}, P_{21}, P_{22} определяется следующими соотношениями:

$$P_{11} = P(A \cap B) = \int_{(D)} \varphi_{\hat{U}/H_0}(U)\, dU, \tag{4.4}$$

$$P_{12} = P(A \cap \overline{B}) = \int_{(Q)} \varphi_{\hat{U}/H_0}(U)\, dU, \tag{4.5}$$

$$P_{21} = P(\overline{A} \cap B) = \int_{(D)} \varphi_{\hat{U}/H_1}(U)\, dU, \tag{4.6}$$

$$P_{22} = P(\overline{A} \cap \overline{B}) = \int_{(Q)} \varphi_{\hat{U}/H_1}(U)\, dU, \tag{4.7}$$

Как видно, из данных выражений, величины вероятностей P_{11}, P_{12}, P_{21}, P_{22} зависят от размеров и расположения областей D и Q. Поэтому, предъявляя соответствующие требования к вероятностям P_{11}, P_{12}, P_{21}, P_{22}, можно определить расположение и границы области допустимых значений и критической области, т. е. критические границы. Практически применение принципов проверки гипотез зависит от объема априорной информации, которая может быть использована при решении задачи. В связи с этим методы задания критической области принято делить на две группы:

- методы, опирающиеся на оценки потерь от неправильного решения;
- методы, опирающиеся на оценки вероятностей ошибок при принятии решения.

Наибольшее распространение получили методы второй группы, так как их применение требует минимальной априорной информации при проверке гипотез, однако, достоверность принятия решений с помощью данных методов ниже, чем для методов первой группы. При проверке гипотез используется четыре вида правил, причем практическое применение правил того или иного вида зависит от полноты априорных данных. Если задача проверки гипотез сформулирована как

задача выбора решений и определена матрица потерь (см. подразд. 3.2), то оптимальное решение может быть получено на основе байесовского или минимаксного правила (подразд. 3.3).

Если функция потерь не определена, для однозначного выбора решений можно использовать два подхода. В случае априорного распределения гипотез наиболее полной характеристикой степени соответствия каждой из них результатом произведенного испытания является апостериорная вероятность этой гипотезы. При этом истинной считается гипотеза, которой соответствует наибольшая апостериорная вероятность. Указанное правило называется правилом апостериорной вероятности.

При отсутствии данных об априорном распределении гипотез единственной характеристикой степени соответствия той или иной гипотезы результатам наблюдений является функция правдоподобия. Поэтому в таких ситуациях выбор решения производится на основе правила максимума правдоподобия, т. е. истинной считается гипотеза, при которой функции я правдоподобия достигает наибольшего значения.

4.4 проверка гипотез классическим методом

Применение рассмотренных методов связано с использованием априорной информации, которая далеко не всегда имеется в расположении исследователя при обработке информации. В связи с этим на практике наибольшее распространение получили методы проверки гипотез, опирающиеся при назначении критических границ только на информацию о вероятностях ошибок первого и второго рода. Сущность данных методов состоит в задании вероятности ошибки первого рода p_{12} с показателем согласованности $\hat{U}$, можно определить границы и расположение критической области Q.

В связи с тем, что вероятность ошибки первого рода играет в данных методах ведущую роль, она имеет специальное название – уровень значимости критерия проверки гипотезы, и специальное обозначение. В дальнейшем она будет обозначаться символом α, хотя в литературе используются и другие обозначения β, q, ξ и т. д. Таким образом, в соответствии с выражением (4.5)

$$\alpha = \mathrm{P}_{12} = \mathrm{P}(\mathrm{A} \cap \overline{\mathrm{B}}) = \int_{(Q)} \varphi_{\hat{U}/H_0}(\mathrm{U})\, d\mathrm{U}, \qquad (4.8)$$

Если известны $\varphi_{\hat{U}/H_0}(\mathrm{U})$ и значение α, то выражение (4.8) позволяет определить критические границы, т. е. границы области Q. Данные границы в дальнейшем будут обозначать символами U_{α_1}, U_{α_2}, …, если область состоит из нескольких подобластей, или символы U_α, если она представляет собой одну область, т. е. возможные значения показателя согласованности $\hat{U}$ делятся границей

U_α на две полупрямые. При конкретном выборе критических границ U_α необходимо учитывать два дополнительных обстоятельства:

- соотношение между условными законами распределения показателя согласованности $\widehat{U}$, соответствующими нулевой и альтернативной гипотезам;

- взаимозависимость ошибок первого и второго рода.

Соотношение между условными законами распределения показателя согласованности $\widehat{U}$, соответствующими нулевой и альтернативной гипотезам, выражается в виде взаимного расположения их кривых распределения на оси абсцисс. Особенности характеристики $\widehat{U}$, а также нулевой и альтернативной гипотез приводят к тому, что кривые распределения $\varphi_{\widehat{U}/H_0}(\mathrm{U})$, $\varphi_{\widehat{U}/H_1}(\mathrm{U})$ могут располагаться друг относительно друга тремя различными способами:

- известно, что кривая распределения $\varphi_{\widehat{U}/H_1}(\mathrm{U})$ сдвинута относительно $\varphi_{\widehat{U}/H_0}(\mathrm{U})$ только вправо (рис. 4.3а);

- известно, что кривая распределения $\varphi_{\widehat{U}/H_1}(\mathrm{U})$ сдвинута относительно $\varphi_{\widehat{U}/H_0}(\mathrm{U})$ влево (рис. 4. 3б);

- сдвиг кривой распределения $\varphi_{\widehat{U}/H_1}(\mathrm{U})$ как вправо, так и влево (рис. 4.3в).

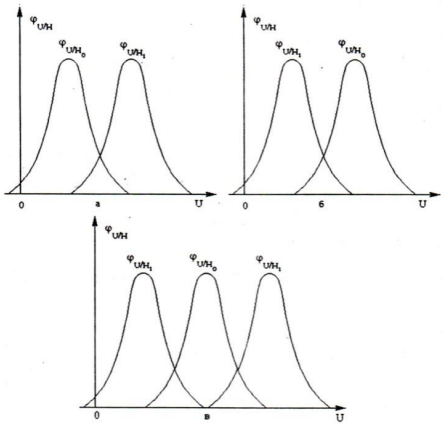


Рис. 4.3

Очевидно, что вид критической области для каждого способа должен быть различен, а именно, в первом случае должна быть выбрана правосторонняя критическая область, во втором – левосторонняя и в третьем – двусторонняя. Правосторонней называют критическую область, определяемую неравенством $\mathrm{U} > U_\alpha$ (рис. 4.4а), левосторонней – неравенством $\mathrm{U} < U_\alpha$ (рис. 4.4б), двусторонней – неравенством $\mathrm{U} < U_\alpha$,

$\mathrm{U} > U_\alpha$, где $U_{\alpha_2} > U_{\alpha_1}$ (рис. 4.4в).

При описании критической области достаточно найти критические точки. Методика их нахождения состоит в следующем. Задаются уровнем зависимости α и ищут критическую точку U_α, исходя из требований, чтобы при условии справедливости нулевой гипотезы вероятность того, что показатель согласованности $\widehat{U}$ попадает в критическую область, была равна принятому уровню значимости.

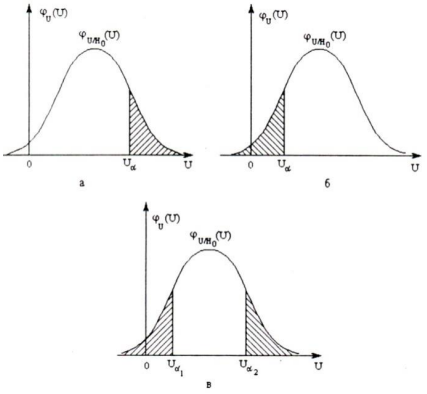


Рис. 4.4

На основании (4.8) для правосторонней критической области это условие имеет вид

$$P(\widehat{U} \geq U_\alpha) = \int_{U_\alpha}^{\infty} \varphi_{\widehat{U}/H_0}(U)\ dU = \alpha, \tag{4.9}$$

для левосторонней

$$P(\widehat{U} < U_\alpha) = \int_{-\infty}^{U_\alpha} \varphi_{\widehat{U}/H_0}(U)\ dU = \alpha, \tag{4.10}$$

Для двусторонней

$$P(\widehat{U} < U_{\alpha_1}) + P(\widehat{U} \geq U_{\alpha_2}) = \int_{-\infty}^{U_1} \varphi_{\widehat{U}/H_0}(U)\ dU + \\ + \int_{U_{\alpha_2}}^{\infty} \varphi_{\widehat{U}/H_0}(U)\ dU = \alpha, \tag{4.11}$$

В последнем случае чаще всего выбирают симметрично расположенные критические точки, т. е.

$$P(\widehat{U} < U_{\alpha_1}) + P(\widehat{U} \geq U_{\alpha_2}) = \frac{\alpha}{2}, \tag{4.12}$$

Критические точки, удовлетворяющие приведенным выше условиям. Находятся по соответствующим таблицам. Таким образом, при выборе критических областей необходимо учитывать не только свойства нулевой, но и свойства альтернативной гипотезы. Для построения взаимозависимости ошибок первого

и второго рода рассмотрим условные плотности $\varphi_{\hat{U}/H_0}(\mathrm{U})$, $\varphi_{\hat{U}/H_1}(\mathrm{U})$ и правостороннюю критическую область критической границей U_α (рис. 4.5).

Из предыдущего известно, что вероятность ошибки первого рода равна вероятности попадания показателя $\hat{U}$ в критическую область, т. е. α. Чем меньше уровень значимости α, тем реже будет допускаться ошибка первого рода, т. е. отвергаться правильная нулевая гипотеза H_0. Однако было бы неверным на основании этого делать вывод о том, что значение вероятности α должно быть выбрано как можно меньшим. Причину этого легко установить, если исходить из предположения о справедливости не гипотезы H_0, а конкурирующей гипотезы H_1. Так как при этом предположении распределение показателя согласованности $\hat{U}$ будет характеризоваться плотностью $\varphi_{\hat{U}/H_1}(\mathrm{U})$, то в соответствии с логикой проверки гипотез будем с вероятностью

$$\beta = \int_{-\infty}^{U_\alpha} \varphi_{\hat{U}/H_1}(\mathrm{U})\ d\mathrm{U} \qquad (4.13)$$

делать заключение о справедливости гипотезы H_0, при условии, что верна гипотеза H_1, т. е. допускать ошибку второго рода. Как видно из рис. 4.5, уменьшение вероятности ошибки первого рода приводит к возрастанию ошибки второго рода, и наоборот.

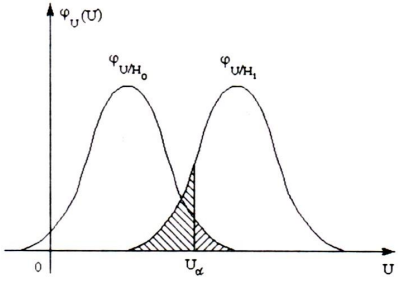


Рис. 4.5

Указанная взаимосвязь ошибок первого и второго рода позволяет сделать следующий важный вывод: для уменьшения вероятности ошибки при принятии гипотезы критическую границу необходимо выбрать таким образом, чтобы сумма вероятностей ошибок первого и второго рода была минимальной (см. подразд. 3.3). Если показатель согласованности подчинен нормальному закону, то минимум суммы вероятностей ошибок первого и второго рода достигается при выборе критической границы в абсциссе точек пересечения кривых распределения $\varphi_{\hat{U}/H_0}(\mathrm{U})$ $\varphi_{\hat{U}/H_1}(\mathrm{U})$ т. е. так, как показано на рис 4.5.

Вместе с тем необходимо заметить, что не во всех случаях подход к выбору U_α с учетом минимума суммы вероятностей ошибок первого и второго рода целесообразен. На практике чаще всего ответ о выборе целесообразно величины α зависит от «тяжести» последствий ошибок первого и второго рода для каждой конкретной задачи. Например, если ошибка первого рода повлечет большие потери, а второго рода – малые, то целесообразно принять возможно меньшее α. Поскольку вероятность ошибки второго рода играет важную роль при выборе критической области, то критерий проверки принято характеризовать так называемой мощностью показателя согласованности. Мощностью χ показателя согласованности называют вероятность попадания показателя согласованности $\hat{U}$ в критическую область при условии, если справедлива конкурирующая гипотеза. Другими словами. Мощность – это вероятность того, что нулевая гипотеза будет отвергнута, если справедлива конкурирующая гипотеза. На основе определения можно записать

$$\chi = \int_{U_\alpha}^{\infty} \varphi_{\hat{U}/H_1}(\mathrm{U})\ d\mathrm{U}, \tag{4.14}$$

С учетом выражения (4.13) равенство (4.14) примет вид

$$\chi = 1 - \int_{-\infty}^{U_\alpha} \varphi_{\hat{U}/H_1}(\mathrm{U})\ d\mathrm{U} = 1 - \beta, \tag{4.15}$$

т. е. мощность показателя согласованности – вероятность того, что не будет допущена ошибка второго рода. Таким образом, для уменьшения ошибки второго рода критическую область необходимо строить так, чтобы мощность показателя согласованности при заданном уровне значимости была максимальной. Мощность показателя согласованности позволяет обоснованно подойти к выбору односторонних критических областей. Предположим, что в качестве критической по-прежнему выбрана правосторонняя область (рис. 4.5), но кривая условной плотности распределения для альтернативной гипотезы $\varphi_{\hat{U}/H_1}(\mathrm{U})$ смещена относительно $\varphi_{\hat{U}/H_0}(\mathrm{U})$ влево (рис.4.6).

Найденные в этих условиях вероятности

$$\beta = \int_{-\infty}^{U_\alpha} \varphi_{\hat{U}/H_1}(\mathrm{U})\ d\mathrm{U} \approx 1; \qquad \chi = 1 - \beta = 0 \tag{4.16}$$

показывают, что при таком подборе критической области показатель согласованности $\hat{U}$ становится непригодным для статистической проверки гипотезы H_0, так как при этом мощность используемого показателя согласованности $\hat{U}$ близка к нулю, а ошибки второго рода становятся практически достоверными. Очевидно, что для увеличения мощности показателя согласованности в данном слу-

чае следует выбрать левостороннюю критическую область. Если характер альтернативной гипотезы неясен, то целесообразно в качестве критической выбрать двустороннюю симметричную область. Заметим, что единственный способ одновременного уменьшения вероятностей ошибок первого и второго рода состоит в увеличении объема выборки.

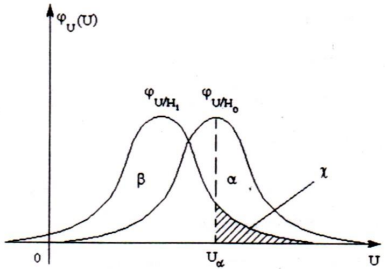


Рис. 4.6

4.5 Методы проверки гипотез о законах распределения.

4.5.1 Постановка задачи.

При обработке результатов наблюдений над случайными объектами одним из основных вопросов является обоснование выбора закона распределения, которому подчинены результаты эксперимента. Пусть экспериментальным путем получена случайная выборка $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$. В связи с ее ограниченностью приходится решать две задачи.

1. Подобрать для получения статистического ряда теоретическую кривую распределения, выражающую лишь существенные черты статистического материала, но не случайности, связанные с недостающим объемом экспериментальных данных (задача выравнивания или сглаживания статистических рядов).

2. Определить, чем объясняются неизбежные расхождения между подобранной теоретической кривой распределения и статистическим распределением: случайными обстоятельствами или тем, что подобранная кривая плохо выравнивает данное статистическое распределение (задача проверки гипотез о законах распределения).

Задача выравнивания статистического распределения заключается в том, чтобы подобрать теоретическую кривую распределения, с той или иной точки зрения наилучшим образом описывающую данное статистическое распределение. Для решения этой задачи, не рассматриваемой в данном учебнике, применяются методы моментов, система кривых К. Пирсона, система кривых Н. А. Бородачева и ряд других методов.

Задача проверки гипотезы о законах распределения начинается с выбора нулевой гипотезы. Для выбора нулевой гипотезы может быть использована следующая методика. По данным эксперимента определяются статистические оценки коэффициента асимметрии $A_{\hat{x}}$ и коэффициента эксцесса $E_{\hat{x}}$:

$$A_{\hat{x}} = \frac{\mu_3[\hat{x}]}{\delta_{\hat{x}}^3}; \qquad E_{\hat{x}} = \frac{\mu_4[\hat{x}]}{\delta_{\hat{x}}^4} - 3, \tag{4.17}$$

где

$$\delta_{\hat{x}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{m}_{\hat{x}})^2}{n-1}} = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - n\hat{m}_{\hat{x}}^2}{n-1}},$$

$$\mu_3[\hat{x}] = \frac{\sum_{i=1}^{n}(x_i - m_{\hat{x}})^3}{n},$$

$$\mu_4[\hat{x}] = \frac{\sum_{i=1}^{n}(x_i - m_{\hat{x}})^4}{n}.$$

В теории распределений доказывается, что каждому закону свойственно определенное соотношение между коэффициентами асимметрии и эксцесса, т. е. может быть построена диаграмма (рис. 4.7). на ней выделены следующие характерные точки, прямые и области. Точки (0; -1,2); (0; 0);

(0; 3); (4; 6) отвечают соответственно равномерному, нормальному, Лапласа и показательному распределению. Так. Для любого нормального закона $A_{\hat{x}} = = 0$, $E_{\hat{x}} = 0$, что и определяют координаты точки II. Гамма-распределение, логарифмически нормальное распределение, распределения Стьюдента и Пуассона показаны на диаграмме прямыми, а бета распределение представлено областью на рис. 4.7 обозначено: I – равномерный закон; II – нормальный закон; III – закон Лапласа; IV – бета-распределение; V – закон Стьюдента; VI – гамма-распределение; VII – закон Пуассона; VIII – показательный закон; IX – логарифмически нормальное распределение.

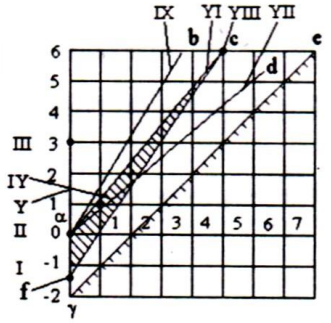


Рис. 4.7

При попадании точки в области диаграммы, для которых не определен закон распределения, выдвижение гипотетического закона должно осуществляться

на основании каких-либо дополнительных априорных соображений. Знание оценок коэффициентов асимметрии и эксцесса позволяет приближенно определить гипотетический закон распределения. Для этого по полученным значениям оценок на диаграмму наносится точка ($A_{\hat{x}}$ $E_{\hat{x}}$). если она окажется вблизи от точки, прямой или области, соответствующих одному из распределений, то последнее и следует выдвинуть в качестве гипотезы. Задачу проверки гипотезы о виде закона распределения рассмотрим в предложении, что предыдущая задача решена, т. е. подобран теоретический закон распределения. Задача проверки гипотезы о законе распределения формулируется следующим образом.

Пусть в результате эксперимента получена случайная выборка $\hat{x}_1$, $\hat{x}_2$, …, $\hat{x}_n$ и для нее подобран теоретический закон распределения, характеризуемый функцией распределения $F_{\hat{x}}(x)$ или плотностью распределения $\varphi_{\hat{x}}(x)$. Необходимо на основании обработки и анализа полученной выборки проверить гипотезу H_0 о том, что наблюдаемая случайная величина подчинена выбранному закону распределения.

В настоящее время существует ряд методов решения данной задачи, однако, на практике наибольшее распространение получили методы А. Н. Колмогорова, К. Пирсона и Н. В. Смирнова, отличающиеся друг от друга видом меры близости между статистическим и гипотетическим законами распределения. Так, в методах А. Н. Колмогорова и Н. В Смирнова такой мерой является функция разности между статистической функцией распределения $F^*_{\hat{x}}(x)$ и функцией распределения $F_{\hat{x}}(x)$ гипотетического закона, т. е.

$$d = f[F^*_{\hat{x}}(x) - F_{\hat{x}}(x)], \tag{4.18}$$

а в методе К. Пирсона – функция разности между относительной частотой и вероятностью попадания случайной величины в заданные интервалы, т. е.

$$d = f[p^*_j - p_j], \tag{4.19}$$

где j – номер интервала.

4.5.2 Проверка гипотез о законе распределения по методу А. Н. Колмогорова

При проверке гипотез о законе распределения по методу А. Н. Колмогорова в качестве показателя согласованности гипотезы используется случайная величина

$$\widehat{U} = \sqrt{n}\ \max_x |F^*_{\hat{x}}(x) - F_{\hat{x}}(x)|, \tag{4.20}$$

где $F^*_{\hat{x}}(x)$ и $F_{\hat{x}}(x)$ – соответственно статистическая и теоретическая (гипотетическая) функции распределения наблюдаемой случайной величины $\hat{x}$. А. Н. Колмогоров доказал, что независимо от вида закона распределения случайной величины $\hat{x}$, функция распределения показателя согласованности $\widehat{U}$ в пределе, т.

е. при n → ∞, характеризуется зависимостью

$$F_{\hat{U}}(\mathrm{U}) = \sum_{k=-\infty}^{\infty}(-1)^k\, e^{-2k^2U^2}\Delta(U).$$

В качестве критической области в методе А. Н. Колмогорова используется область. Следовательно, из равенства

$$\alpha = \mathrm{p}(\hat{U} \geq U_\alpha) = 1 - \mathrm{p}(\hat{U} < U_\alpha) = 1 - F_{\hat{U}}(\mathrm{U})$$

вытекает, что критическая граница U_α равняется квантили случайной величины $\hat{U}$ при аргументе $1 - \alpha$:

$$U_\alpha = F_{\hat{U}}^{-1}(1-\alpha). \qquad (4.21)$$

Для значений U_α, рассчитанных по формуле (4.21). составлена специальная таблица (4.1):

Таблица 4.1

α	0,5	0,1	0,05	0,01	0,001
U_α	0,828	1,224	1,358	1,627	1,950

Общий вид зависимости (4.21) изображен на рис. 4.8.

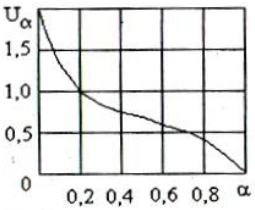


Рис. 4.8

Правило проверки гипотезы о законе распределения заключается в следующем:

Назначается уровень зависимости α и по табл. 4.1 или по графику на рис. 4.8 определяется критическая граница U_α.

По результатам наблюдений строится статистическая функция распределения $F_{\hat{x}}^*(\mathrm{x})$, показанная на рис. 4.9.

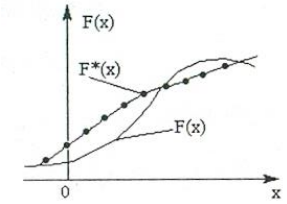


Рис. 4.9

На том же графике строится предполагаемая теоретическая функция распределения $F_{\hat{x}}(\mathrm{x})$.

По графику определяется максимальная величина модуля разности ординат статистической и теоретической функции распределения и вычисляется значение U по формуле (4.20).

Проверяется условие $U > U_\alpha$. Если оно выполняется, то гипотеза H_0 бракуется, в противном случае делается вывод, что результаты эксперимента не противоречат гипотезе о том, что наблюдаемая случайная величина $\hat{x}$ подчинена закону распределения с функцией распределения $F_{\hat{x}}(x)$.

Достоинствами метода А. Н. Колмогорова являются его простота и отсутствие сложных расчетов. Однако он обладает следующими существенными недостатками:

применение метода требует значительной априорной информации и о гипотетическом законе распределения, так как кроме вида закона распределения должны быть указаны значения всех параметров распределения;

метод учитывает только максимальное отклонение статистической функции распределения от теоретической, а не закон изменения этого отклонения по всему размаху случайной выборки.

В связи с этим при принятии гипотезы может быть допущена ошибка в тех случаях, когда функция распределения сдвинута по оси абсцисс (рис. 4.10).

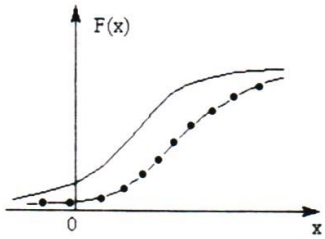


Рис. 4.10

Так если соотношение между теоретической и статистической функциями распределения имеет вид, изображенный на рис. 4.9, причем значения $\max_x |F^*_{\hat{x}}(x) - F_{\hat{x}}(x)|$ на рис. 4.9 и 4.10 имеют приблизительно одну и ту же величину, то при использовании метода А. Н. Колмогорова будут приняты одинаковые выводы, хотя во втором случае (см. рис. 4.10) соответствие гипотезы опытным данным значительно хуже, чем в первом.

4.5.3 Проверка гипотез о законе распределения по методу Н. В. Смирнова

При поверке гипотезы о законе распределения по методу Н. В. Смирнова (методу w^2) в качестве меры близости теоретического и статистического законов распределения, как и в методе А. Н. Колмогорова, используется функция разности статистической и теоретической функции распределения. Однако в качестве

показателя согласованности гипотезы применяется не максимальное значение этой разности, а среднее ее значение по всей области определения функции распределения, что исключает недостаток, присущий методу А. Н. Колмогорова.

В общем случае показатель согласованности гипотезы определяется выражением

$$\widehat{U} = \widehat{w}^2 = (\mathrm{s}) \int_{-\infty}^{\infty} [F_{\hat{x}}^*(\mathrm{x}) - F_{\hat{x}}(\mathrm{x})]^2 \, dF_x(\mathrm{x}), \tag{4.22}$$

так как наблюдаемая случайная величина может быть смешанной. Здесь (s) – символ интеграла Стильтьеса.

В частном случае для непрерывной случайной величины

$$dF_{\hat{x}}(\mathrm{x}) = \varphi_{\hat{x}}(\mathrm{x}) d\mathrm{x}$$

и интеграл Стильтьеса переходит в обычный интеграл Римана:

$$\widehat{U} = \widehat{w}^2 = \int_{-\infty}^{\infty} [F_{\hat{x}}^*(\mathrm{x}) - F_{\hat{x}}(\mathrm{x})]^2 \, \varphi_{\hat{x}}(\mathrm{x}) d\mathrm{x}, \tag{4.23}$$

Если наблюдаемая случайная величина является дискретной, то интеграл Стильтьеса переходит в сумму

$$\widehat{U} = \widehat{w}^2 = \sum_{i=1}^{n} [F_{\hat{x}}^*(\mathrm{x_i}) - F_{\hat{x}}(\mathrm{x_i})]^2 \, p_i, \tag{4.24}$$

Можно показать, что после ряда преобразований выражение (4.24) приводится к виду

$$\widehat{U} = \frac{1}{12n^2} + \frac{1}{n} \sum_{i=1}^{n} [F_{\hat{x}}^*(\mathrm{x_i}) - \frac{2i-1}{2n}]^2, \tag{4.25}$$

применение которого значительно упрощает процесс вычисления показателя согласованности.

Точный закон распределения случайной величины $\widehat{U}$, определяемой выражением (4.25), имеет весьма сложный вид, поэтому на практике в качестве согласованности гипотезы используют случайную величину

$$\widehat{U} = \frac{1}{12n} + \sum_{i=1}^{n} [F_{\hat{x}}^*(\mathrm{x_i}) - \frac{2i-1}{2n}]^2, \tag{4.26}$$

закон распределения которой уже при n = 40 достаточно близок к предельному закону, график функции распределения которого изображен на рис. 4.11.

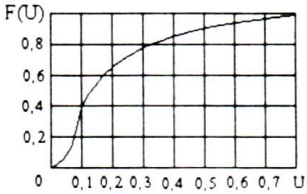


Рис.4.11

Значения критической границы U_α для этого предельного закона, равные

квантилям случайной величины $\hat{U}$ порядка (1 – a) и рассчитать по формуле 4.21) в зависимости от уровня значимости a, приведены в табл. 4.2:

Таблица 4.2

a	0,5	0,1	0,05	0,02	0,01
U_α	0,118	1,347	1,461	1,620	1,744

Проверка гипотезы значимости a и по табл. 4.2 определяется нижняя граница U_α критической области.

Вычисляется наблюдаемое значение показателя согласованности U по формуле (4.26).

Проверяется условие U $>$ U_α. Если оно выполняется, то нулевая гипотеза отвергается, в противном случае делается вывод о том, что результаты наблюдений не противоречат гипотезе о подчинении наблюдаемой случайной величины $\hat{x}$ теоретическому закону распределения с функцией распределения $F_{\hat{x}}(x)$.

4.5.4 Проверка гипотез о законе распределения по методу К. Пирсона

В методе К. Пирсона мерой близости теоретического и статистического законов распределения служит сумма квадратов разностей между относительной частотой и вероятностью попадания случайной величины x в интервалы, на которые разбивается множество возможных значений этой величины:

$$\hat{U} = \sum_{j=1}^{r} c_j(\hat{p}_j^* - p_j)^2, \qquad (4.27)$$

где r – число интервалов;

j – номер интервала.

Коэффициенты c_j вводятся для учета обстоятельства, что абсолютные того значения разностей $\hat{p}_j^* - p_j$ неравнозначны при различных значениях p_j.

Действительно, одно и то же значение разности $\hat{p}_j^* - p_j$ является малозначимым при большой величине p_j и представляет собой заметную величину, если вероятность p_j мала. К. Пирсон показал, что коэффициент c_j целесообразно брать обратно пропорциональными вероятностям p_j, причем, если определять их на основе выражения

$$c_j = \frac{n}{p_j}, \quad j = \overline{1.r}, \qquad (4.28)$$

то при больших значениях n закон распределения случайной величины

$$\hat{U} = \sum_{j=1}^{r} \frac{n(\hat{p}_j^* - p_j)^2}{p_j} \qquad (4.29)$$

обладает весьма важным свойством: он практически зависит не от вида закона распределения случайной величины $\hat{x}$ и объема выборки n, а только от

числа интервалов r, причем при увеличении n закон распределения случайной величины $\widehat{U}$ приближается к распределению χ^2. Докажем это утверждение.

Рассмотрим случайную величину $\widehat{m}_j$ - число попаданий наблюдаемой случайной величины $\hat{x}$ в j-тый интервал. Эта случайная величина распределена по биноминальному закону распределения с характеристиками:

$$M[\widehat{m}_j] = np_j;\ \ \delta[\widehat{m}_j] = \sqrt{np_j(1-\ p_j)}. \tag{4.30}$$

Однако при достаточно большом n величину $\widehat{m}_j$ на основании теоремы Муавра-Лапласа можно считать распределенной по нормальному закону с теми же числовыми характеристиками. Прономеровав случайную величину $\widehat{m}_j$, получим

$$\hat{z}_j\ = \frac{\widehat{m}_j - np_j}{\sqrt{np_j(1-p_j)}}. \tag{4.31}$$

Нормированные случайные величины $\hat{z}_j$ связаны между собой линейным соотношением

$$\sum_{j=1}^{r} \hat{z}_j\ \sqrt{np_j(1-p_j)} = \sum_{j=1}^{r} \widehat{m}_j\ - n \sum_{j=1}^{r} p_j\ = n - n\ = 0.$$

На основании этого случайная величина $\sum_{j=1}^{r} \hat{z}_j{}^2$ будет приблизительно следовать хи-квадрат распределению с (r – 1) степенями свободы. Если эту случайную величину принять за статистическую характеристику гипотезы, то получим равенство

$$\widehat{U} = \chi^2_{(r-1)} = \sum_{j=1}^{r} \frac{(\widehat{m}_j - np_j)^2}{np_j(1-p_j)}. \tag{4.32}$$

Преобразуем выражение (4.32), учитывая, что

$$\sum_{j=1}^{r} m_j\ = n$$

и $1 - p_j \approx 1$ при больших значениях n:

$$\widehat{U} = \sum_{j=1}^{r} \frac{n^2\left(\frac{\widehat{m}_j}{n} - p_j\right)^2}{np_j\,(1-p_j)} = \sum_{j=1}^{r} \frac{n(\hat{p}_j^* - p_j)^2}{p_j} \tag{4.33}$$

или

$$\widehat{U} = \sum_{j=1}^{r} \frac{\widehat{m}_j{}^2 - 2\widehat{m}_j np_j + n^2p_j^2}{np_j} = \sum_{j=1}^{r} \frac{\widehat{m}_j{}^2}{np_j} - 2\sum_{j=1}^{r} \widehat{m}_j + \\ + n\sum_{j=1}^{r} p_j\ = \sum_{j=1}^{r} \frac{\widehat{m}_j{}^2}{np_j}\ -\ n. \tag{4.34}$$

Выражения (4.33) или (4.34) используются в зависимости от формы пред-

ставления результатов наблюдения, т. е. в зависимости от того, являются ли исходными данными $\hat{p}_j^*$ или $\hat{m}_j$. как известно, распределение хи-квадрат зависит от числа степеней свободы k = r – s, равного числу интервалов r минус число независимых условий («связей»), наложенных на частоты $\hat{p}_j^*$. В формуле (5.33) предполагается наличие только одной связи (s = 1), представляющей собой условие

$$\sum_{j=1}^{r} \hat{p}_j^* = 1, \qquad (4.35)$$

которое накладывается во всех случаях.

В случае, когда теоретическое распределение подбирается так, чтобы совпадали математическое ожидание теоретического распределения и оценка математического ожидания, полученная по результатам наблюдения, т. е.

$$\sum_{j=1}^{r} \overline{x_l} p_j^* = M_{\hat{x}}, \qquad (4.36)$$

число связей увеличивается на единицу. Следовательно, s = 2 и число степеней свободы k = r – 2. Если условие совпадения параметров теоретического закона распределения и статистического распределения распространяется и на дисперсию, т. е. предполагается, что

$$\sum_{j=1}^{r} (\overline{x_l} - M_{\hat{x}})^2 \, p_j^* = D_{\hat{x}}, \qquad (4.37)$$

то s = 3 и k = r – 3, и т. д.

таким образом, число степеней свободы хи-квадрат распределения при проверке гипотез зависит от условий проведения проверки, что необходимо учитывать, используя показатель согласованности гипотезы (4.33) или (4.34). можно показать, что при невыполнении гипотезы H_0 по мере возрастания n значение показателя согласованности $\hat{U}$ будет неограниченно увеличиваться, т. е. кривая распределения $\varphi_{\hat{U}/H_1}(U)$ сдвинута относительно кривой $\varphi_{\hat{U}/H_0}(U)$ вправо. Поэтому в качестве критической целесообразно выбрать правостороннюю критическую область (рис. 4.12). В этом случае для определения критической границы U_a можно использовать таблицы, в которых даны значения 100%-ного предела для закона хи-квадрат в зависимости от вероятности a и числа степеней свободы k. (Таблицы приведены в приложении).

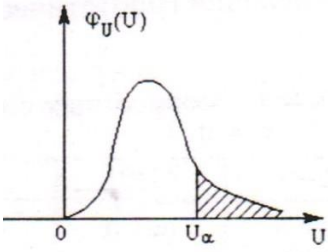


Рис. 4.12

Порядок проверки гипотезы о виде закона распределения состоит в следующем:

Назначается уровень значимости a, и по табл. 3 (приложение) определяется критическая граница U_a. Входами в таблицу служат уровень значимости a и число степеней свободы k.

Результаты эксперимента представляются в виде статистического ряда (табл. 4.3), где m_j и p_j^* - частота и частость попадания исследуемой величины $\hat{x}$ в j-тый интервал соответственно.

Таблица 4.3

I = $[x_j^r, p_{j+1}^r]$	x_1^r, x_2^r	x_1^r, x_2^r	…	x_j^r, x_{j+1}^r	…	x_r^r, x_{r+1}^r
m_j	m_1	m_2	…	m_j	…	m_r
p_j	p_1	p_2	…	p_j^*	…	p_r^*

Вычисляются вероятности p_j попадания случайной величины $\hat{x}$, подчиняющейся гипотетическому закону распределения, в j-тый интервал

(j = $\overline{1,r}$):

$$p_j = p\,(x_j^r < \hat{x} < x_{j+1}^r = \int_{x_j^r}^{x_{j+1}^r} \varphi_{\hat{x}^r}(x)dx \qquad (4.38)$$

где $\varphi_{\hat{x}^r}(x)$ – плотность распределения гипотетического закона. Очевидно, что

$\sum_{j-1}^{r} p_j = 1$, как сумма вероятностей несовместных событий, образующих полную группу.

Рассчитывается значение U показателя согласованности гипотезы по формуле (4.33) или (4.34).

Проверяется условие $U \leq U_a$. Если оно выполняется, то расхождение между экспериментальными данными и гипотезой H_0 полагается несущественными. В противном случае нулевая гипотеза отвергается.

Существенное достоинство метода К. Пирсона состоит в возможности его

применения тогда, когда априорно известен лишь вид гипотетического распределения, но не известны его параметры. В этом случае параметры распределения заменяются оценками, полученными по экспериментальным данным и используемыми в дальнейшем для вычисления вероятностей p_j, а число степеней свободы уменьшается на число заменяемых параметров. Метод К. Пирсона имеет следующие недостатки:

он применим только при большой выборке (n $\geq$ 100), так как показатель согласованности подчиняется распределению хи-квадрат лишь при достаточно большом n;

результаты проверки в значительной степени зависят от способа разбиения выборки на интервалы, причем их число может быть не менее 10, а количество попаданий случайной величины $\hat{x}$ в любой из интервалов – не менее 5.

Пример 2. Проверить, пользуясь критерием согласия Пирсона χ^2, соответствие статистического ряда распределения температуры воздуха T, построенного по результатам 100 наблюдений над ней, нормальному закону. Уровень значимости a принять равным 0,05.

Таблица 4.4

Характеристика	Градации, °C				
	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30
p_i^*	0.070	0.190	0.380	0.240	0.120
p_i	0.063	0.211	0.352	0.268	0.106

На основе анализа статистического распределения температуры воздуха T предполагаем, что она имеет нормальное распределение, т. е.

$$f(t) = \frac{1}{\delta_t \sqrt{2\pi}} e^{-(t-m_t)^2/2\delta_t^2}.$$

Найдем по данным статистического ряда распределения оценки параметров выбранной выравнивающей функции m_t и δ_t:

$$\bar{t} = \sum_{i=1}^{6} t_i p_i^* = 7{,}5*0{,}07 + 12{,}5*0{,}19 + \ldots + 27{,}5*0{,}12 \approx 18{,}25\ ^\circ\text{C};$$

$$D_t^* = \sum_{i=1}^{6} t_i^2 p_i^* - \bar{t}^2 = 7{,}5^2\ 0{,}07 + 12{,}5^2\ 0{,}19 + \ldots + 27{,}5^2\ 0{,}12 - 18{,}25^2 \approx 29\ ^\circ\text{C}^2$$

и значит, $s_t^* = \sqrt{D_t^*} = 5{,}38^\circ\text{C}$. Принимаем: $m_t = t = 18{,}25^\circ\text{C}$ и $\delta_t = s_t^* = 5{,}38^\circ\text{C}$.

Тогда искомая выравнивающая функция выразится формулой

$$f(t) = \frac{1}{\delta_t \sqrt{2\pi}} e^{-(t-m_t)^2/2\delta_t^2}.$$

Проверим, пользуясь критерием χ^2, гипотезу H_0, состоящую в том, что распределение рассматриваемой температуры воздуха нормальное с параметрами $m_1 = 18{,}25^\circ\text{C}$ и $\delta_t = 5{,}38^\circ\text{C}$. Для этого:

-находим число степеней свободы

$$r = k - 1 - 1 = 5 - 2 - 1 = 2;$$

по известным значениям α и r из табл. 3 (приложение) находим $\chi^2_{кр} = 5{,}99$;

- с использованием таблицы функции Лапласа вычисляем вероятности p_i по формулам:

$$p_i = F(x_{i+1}) - F(x_i), \qquad (i = 2, 3, 4),$$

$$p_1 = F(x_2),\ p_6 = 1 - F(x_4);$$

например, $p_1 = \Phi\left(\frac{10-18{,}25}{5{,}38}\right) + 0{,}5 = -0{,}437 + 0{,}5 = 0{,}063,$

$$p_2 = \Phi\left(\frac{15-18{,}25}{5{,}38}\right) - \Phi\left(\frac{10-18{,}25}{5{,}38}\right) = 0{,}211,$$

$p_6 = 1 - \Phi\left(\frac{25-18{,}25}{5{,}38}\right) - 0{,}5 = 0{,}106$, и заносим их в таблицу в начале примера; по данным этой таблицы определяем выборочное значение критерия χ^2:

$$\chi^2_H = n\sum_{i=1}^{6} \frac{(p_i^* - p_i)^2}{p_i} = 100\left[\frac{(0.070-0.063)^2}{0.056} + \cdots + \frac{(0.120-0.106)^2}{0.116}\right] = 1.62.$$

Получим, что $\chi^2_n < \chi^2_{кр}$, поэтому выдвинутую гипотезу H_0 можно считать правдоподобной.

4. 6 Методы проверки гипотез о параметрах законов распределения

4.6.1 Проверка гипотез о равенстве математических ожиданий

Пусть имеется две независимые случайные величины $\hat{x}$ и $\hat{y}$, распределенные по нормальному закону. Эксперимент состоит в том, что над случайными величинами $\hat{x}$ и $\hat{y}$ осуществляется соответственно n и m независимых наблюдений, т. е. получаются случайные выборки $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ и $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. По этим выборкам определяются оценки математических ожиданий

$$\widehat{M}_{\hat{x}} = \frac{1}{n}\sum^{n} \hat{x}_i; \qquad \widehat{M}_{\hat{y}} = \frac{1}{m}\sum_{j=1}^{n} \hat{y}_j. \qquad (4.39)$$

Требуется по полученным оценкам проверить гипотезу о равенстве математических ожиданий $\widehat{M}_{\hat{x}}$ и $\widehat{M}_{\hat{y}}$. Такая задача ставится потому, что, как правило, оценки математических ожиданий оказывается различными. Причина этого может быть двоякой: либо действительно отличны и оценки, и математические ожидания $\widehat{M}_{\hat{x}}$ и $\widehat{M}_{\hat{y}}$, либо $\widehat{M}_{\hat{x}}$ и $\widehat{M}_{\hat{y}}$ одинаковы, а отличие оценок вызвано случайными причинами, в частности, случайным отбором вариантов выборки. Если окажется, что нулевая гипотеза справедлива, т. е. $\widehat{M}_{\hat{x}}$ и $\widehat{M}_{\hat{y}}$ одинаковы, то различие в оценках $\widehat{M}_{\hat{x}}$ и $\widehat{M}_{\hat{y}}$ будет обусловлено случайными причинами. В противном случае это различие не может быть объяснено случайными причинами, а определяется отличием математических ожиданий. При решении данной задачи могут встретиться два случая, а именно: случай, когда дисперсии $\widehat{D}_{\hat{x}}$ и $\widehat{D}_{\hat{y}}$ известны,

и случай, когда они не известны.

Рассмотрим первый случай, т. е. осуществим проверку гипотез при известных дисперсиях.

В качестве показателя согласованности гипотезы выберем случайную величину

$$\hat{U} = \frac{\hat{M}_{\hat{x}} - \hat{M}_{\hat{y}}}{\delta(\hat{M}_{\hat{x}} - \hat{M}_{\hat{y}})}. \quad (4.40)$$

Целесообразность выбора такого показателя согласованности определяется следующими соображениями.

Введем в рассмотрение случайную величину

$$\hat{z} = \hat{M}_{\hat{x}} - \hat{M}_{\hat{y}}. \quad (4.41)$$

которая, очевидно, распределена по нормальному закону и имеет числовые характеристики:

$$M_{\hat{z}} = M_{\hat{x}} - M_{\hat{y}}; \qquad D_{\hat{z}} = \frac{\delta_{\hat{x}}^2}{n} + \frac{\delta_{\hat{y}}^2}{m}; \qquad \delta_{\hat{z}} = \sqrt{\frac{\delta_{\hat{x}}^2}{n} + \frac{\delta_{\hat{y}}^2}{m}}.$$

Проанализировав случайную величину $\hat{z}$, получим

$$\hat{U} = \frac{\hat{z}}{\delta_z} = \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i - \frac{1}{m}\sum_{j=1}^{n} \hat{y}_j}{\sqrt{\frac{\delta_{\hat{x}}^2}{n} + \frac{\delta_{\hat{y}}^2}{m}}}. \quad (4.42)$$

Случайная величина $\hat{U}$ подчинена нормальному закону распределения, параметры которого известны: $M_{\hat{U}} = 0$, $\delta_{\hat{U}} = 1$, что существенно упрощает процедуру проверки нулевой гипотезы. Действительно, если гипотеза H_0 справедлива, т. е. $M_{\hat{x}} = M_{\hat{y}}$, то случайная величина $\hat{z}$центрирована, откуда следует, что $M_{\hat{U}} = 0$. Так как выборки независимые, то $\delta_{\hat{U}} = 1$. Критическая область строится в зависимости от вида альтернативной гипотезы, которая может быть сформулирована тремя различными способами:

$$M_{\hat{x}} \neq M_{\hat{y}}; \; M_{\hat{x}} > M_{\hat{y}}; \; M_{\hat{x}} < M_{\hat{y}}.$$

Рассмотрим методику проверки гипотезы H_0 для каждого из приведенных способов формулировки конкурирующей гипотезы.

$$H_0: M_{\hat{x}} = M_{\hat{y}}; \; H_1: M_{\hat{x}} \neq M_{\hat{y}};$$

В этом случае строят двустороннюю критическую область, исходя из требования, чтобы вероятность попадания в нее показателя согласованности при предположении о справедливости нулевой гипотезы была равна принятому уровню значимости a.

Наибольшая мощность критерия достигается тогда, когда левая и правая критическая точка U_{a_1}, U_{a_2} выбраны так, что вероятность попадания показателя согласованности U в каждый из двух интервалов критической области равна $a/2$,

т. е.

$$\left.\begin{aligned} p(\widehat{U} < U_{a_1}) &= \frac{\alpha}{2}, \\ p(\widehat{U} \geq U_{a_2}) &= \frac{\alpha}{2}. \end{aligned}\right\}$$

Поскольку U – нормированная нормально распределенная случайная величина и ее распределение симметрично относительно нуля, то критические точки тоже симметричны относительно нуля, т. е.

$$|U_{a_1}| = |U_{a_2}| = |U_\alpha|.$$

Используя функцию Лапласа, вероятность попадания показателя согласованности в критическую область можно определить выражением

$$1 - \mathrm{p}(|U| < U_\alpha) = 1 - \Phi_1\left(\frac{U_\alpha}{\sqrt{2}}\right) = \alpha,$$

откуда

$$U_a = \sqrt{2}\ \Phi_3^{-1}(1 - a) = t_{1-\alpha} = t_\gamma. \qquad (4.43)$$

Функция $t_\gamma = \sqrt{2}\Phi_1^{-1}(1 - \alpha)$ табулирована табл.2 (приложение), что позволяет находить величину U_a, не пользуясь таблицами функции Лапласа. Заметим, что в таблицах используется значение $\gamma = 1 - \alpha$ (см. (4.43)). Двусторонняя критическая область будет определяться неравенствами $U < -U_\alpha$, $U > U_\alpha$. Таким образом, правило проверки гипотезы H_0 для рассматриваемого случая можно сформулировать следующим образом:

а) назначается уровень значимости α и с помощью табл. 2 (приложение) и формулы (4.43) определяются границы критической области $U_{a_1} = -U_\alpha$; $U_{a_2} = U_\alpha$;

б) на основании результатов выборки вычисляется наблюдаемое значение показателя U по формуле (4.42);

в) проверяется условие $|U| > U_\alpha$. Если оно выполняется, то гипотеза H_0 отвергается. В противном случае данные эксперимента не противоречат нулевой гипотезе и ее следует принять.

Пример 3. На двух расположенных рядом метеостанциях получены ряды наблюдений за температурой, причем на первой станции ряд составляет 10 лет, а на второй – 15. В результате расчетов получены следующие среднеарифметические отклонения температуры от среднегодовых температур для данного региона: для первой станции отклонение равно -0,8℃, для второй +0,4℃. Среднеквадратические отклонения температуры для этих метеостанций известны и равны соответственно 2 и 1,5℃. Необходимо проверить гипотезу о равенстве матема-

тических ожиданий температур на обеих станциях. Обозначим отклонение температур от среднегодовой для заданного региона для первой и второй станций соответственно символами $\hat{x}$ и $\hat{y}$. По условию задачи n = 10, m = 15, $M_{\hat{x}}$ = -0,8°, $M_{\hat{y}}$ = +0.4°, $\delta_{\hat{x}}$ = 2°, $\delta_{\hat{y}}$ = 1,5°. Задаемся уровнем значимости α = 0,05 и по таблице находим

$$U_{\alpha=t_{1-\alpha}} = t_{\gamma=t_{0,96}} = 1{,}96.$$

Вычислив значение U по формуле (4.42), получим

$$|U| = \left| \frac{-0{,}8-0{,}4}{\sqrt{\frac{4}{10}+\sqrt{\frac{2{,}25}{15}}}} \right| = 1{,}62.$$

Так как $|U| < U_{\alpha}$, то нулевая гипотеза $M_{\hat{x}} = M_{\hat{y}}$ не противоречит данным расчетов. Следовательно, есть основание полагать, что среднегодовые температуры обеих станций совпадают.

2. $H_0: M_{\hat{x}} = M_{\hat{y}}$; $H_1: M_{\hat{x}} > M_{\hat{y}}$.

На практике такой случай возможен, если априорные сведения позволяют предложить, что математическое ожидание случайной величины $\hat{x}$ больше математического ожидания случайной величины $\hat{y}$. например, если введено усовершенствование технологического процесса, то естественно допустить, что оно приведет к увеличению выпуска продукции, т. е. математическое ожидание объема выпущенной продукции будет больше. В этом случае строят правостороннюю критическую область такую, чтобы вероятность попадания в нее показателя согласованности при предположении о справедливости нулевой гипотезы была равна принятому уровню значимости α, т. е.

$$P(\hat{U} \geq U_{\alpha}) = \alpha. \qquad (4.44)$$

Покажем, как найти критическую точку с помощью функции Лапласа, перепишем выражение (4.44) в виде:

$$p(\hat{U} \geq U_{\alpha}) = p(U_{\alpha} < \hat{U} < \infty) = \frac{1}{2}\left[1 - \Phi_1\left(\frac{U_{\alpha}}{\sqrt{2}}\right)\right] = \alpha,$$

отсюда получим

$$\frac{1}{2}\Phi_1\left(\frac{U_{\alpha}}{\sqrt{2}}\right) = \frac{1}{2} - \alpha$$

и, следовательно,

$$\Phi_1\left(\frac{U_{\alpha}}{\sqrt{2}}\right) = 1 - 2\alpha.$$

Таким образом,

$$U_\alpha = \sqrt{2}\Phi_1^{-1}(1-2\alpha) = t_{1-2\alpha} \qquad (4.45)$$

Правило проверки для рассматриваемого случая формулируется следующим образом:

а) назначается уровень значимости α и по табл. 2 (приложение), входя в нее со значением (1 - 2α), определяется величина U_α;

б) на основании результатов выборок по формуле (4.42) определяется величина U;

в) проверяется условие $U > U_\alpha$. Если оно выполняется, то гипотеза H_0 отвергается, в противном случае она принимается

3. $H_0: M_{\hat{x}} = M_{\hat{y}}$; $H_1: M_{\hat{x}} < M_{\hat{y}}$.

При такой формулировке альтернативной гипотезы строят левостороннюю критическую область, такую, чтобы вероятность попадания в нее показателя согласованности при предложении о справедливости нулевой гипотезы была равна принятому уровню значимости α, т. е.

$$p(\hat{U} < U_{\alpha_1}) = \alpha.$$

Учитывая, что показатель $\hat{U}$ имеет симметричное распределение относительно нуля, заключаем, что U_{α_1} точка симметрична такой точке $U_\alpha > 0$, для которой $p(\hat{U} \geq U_\alpha) = \alpha$, т. е. $U_\alpha = -U_{\alpha_1}$. В связи с этим методика определения U_{α_1} полностью совпадает с методикой предыдущего случая, только полученное значение берется с отрицательным знаком. Правило проверки гипотезы также аналогично рассмотренному выше правилу. За исключением п. «в», а именно, если $U < U_\alpha$, то нулевая гипотеза отвергается, если $U > U_\alpha$, то принимается. Выше предполагалось, что случайные величины $\hat{y}$ и $\hat{x}$ распределены нормально, а их дисперсии известны. При этих предположениях показатель согласованности гипотезы распределен по нормальному закону с параметрами $M_{\hat{U}} = 0$, $\delta_{\hat{U}} = 1$. Если хотя бы одно из приведенных предположений не выполняется, описанный механизм проверки гипотезы о равенстве математических ожиданий не применим. Однако, если независимые выборки имеют большой объем (не менее 30 вариантов каждая), то оценки математических ожиданий и дисперсий распределены приближенно нормально и закон распределения $\hat{U}$ можно считать близким к нормальному. При этом в качестве показателя согласованности можно использовать случайную величину $\hat{U}^*$ аналогичного вида и проводить проверку гипотезы по описанной выше методике, но к полученным выводам следует относиться с осторожностью.

В случае, когда дисперсии $D_{\hat{x}}$ и $D_{\hat{y}}$ неизвестны, решать задачу проверки гипотезы описанным выше методом нельзя, так как показатель согласованности

вида (4.40) не будет подчиняться нормальному закону распределения. В этом случае задачу проверки гипотезы о равенстве математических ожиданий можно решить лишь при наличии дополнительных данных, а именно: если известно отношение дисперсий

$$\frac{D_{\hat{y}}}{D_{\hat{x}}} = \frac{\delta_{\hat{y}}^2}{\delta_{\hat{x}}^2} = \alpha. \qquad (4.46)$$

В частном случае дисперсии $D_{\hat{x}}$ и $D_{\hat{y}}$ могут быть равны, т. е. $\alpha = 1$. Если отношение (4.46) известно, то можно показать, что наиболее целесообразен показатель согласованности

$$\widehat{U} = \hat{t}_{(n+m-2)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - \frac{1}{m}\sum_{j=1}^{m}\hat{y}_j}{\sqrt{\alpha\sum_{i=1}^{n}(\hat{x}_i - \hat{m}_{\hat{x}})^2 + \sum_{j=1}^{m}(\hat{y}_j - \hat{m}_{\hat{y}})^2}}\sqrt{\frac{amn(n+m-2)}{m+an}}, \qquad (4.47)$$

подчиняющийся закону распределения Стьюдента с (n + m – 2) степенями свободы.

Выражение (4.47) может быть получено на основе равенства (4.40), если в качестве $M_{\hat{x}}$ и $M_{\hat{y}}$ использовать из подходящие оценки

$$\widehat{\widehat{M}}_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i; \qquad \widehat{\widehat{M}}_{\hat{y}} = \frac{1}{m}\sum_{j=1}^{m}\hat{y}_j, \qquad (4.48)$$

а в качестве дисперсий случайных величин $\hat{x}$ и $\hat{y}$ - оценки

$$\hat{\hat{\delta}}_{\hat{x}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{x}_i - \widehat{\widehat{m}}_{\hat{x}})^2; \qquad \hat{\hat{\delta}}_{\hat{y}}^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{y}_j - \widehat{\widehat{m}}_{\hat{y}})^2. \qquad (4.49)$$

Проведением несложных преобразований выражение (4.47) приводится к виду

$$\widehat{U} = \frac{\widehat{\widehat{M}}_{\hat{x}} - \widehat{\widehat{M}}_{\hat{y}}}{\sqrt{a(n-1)\hat{\hat{\delta}}_{\hat{x}}^2 - (m-1)\hat{\hat{\delta}}_{\hat{y}}^2}}. \qquad (4.50)$$

Особенность данного показателя согласованности состоит в том, что его использование не требует априорной информации.

Критическая область, как и ранее, в зависимости от альтернативной гипотезы строится по-разному, а именно:

1. $H_0: M_{\hat{x}} = M_{\hat{y}}; \quad H_1: M_{\hat{x}} \neq M_{\hat{y}}.$

В этом случае строят двустороннюю критическую область, причем в связи с тем, что кривая распределения Стьюдента симметрична относительно нуля, критическая область является симметричной. Ее границы U_α можно найти из равенства

$$2\int_0^{U_\alpha}\varphi_{\hat{t}_{(n+m-2)}}(t)dt = 1 - \alpha \qquad (4.51)$$

с помощью таблицы распределения Стьюдента табл. 6 (приложение).

Правило проверки гипотезы о равенстве математических ожиданий формулируется следующим образом:

а) назначается уровень значимости α и по табл. 6 (приложение) находится величина $U_\alpha = t_{\alpha,k}$. Входами в таблицу являются вероятность

$(1 - \alpha)$ и число степеней свободы k = n + m -2;

б) вычисляется наблюдаемое значение показателя согласованности U по формуле (4.47) или (4.60);

в) проверяется условие $|U| > U_\alpha$. Если оно выполняется, то гипотеза H отвергается. При $|U| < U_\alpha$ данные эксперимента не противоречат нулевой гипотезе и она принимается

2. $H_0: M_{\hat{x}} = M_{\hat{y}}; \quad H_1: M_{\hat{x}} > M_{\hat{y}}.$

В этом случае строят правостороннюю критическую область таким образом, чтобы выполнялось условие

$$p(\hat{U} \geq U_\alpha) = \alpha.$$

Границу критической области можно найти из равенства

$$\int_{U_\alpha}^{\infty} \varphi_{\hat{t}_{(n+m-2)}}(t)dt = \alpha \qquad (4.52)$$

По специальным таблицам распределение Стьюдента.

При выполнении неравенства $U < U_\alpha$ нулевая гипотеза принимается, в противном случае она отвергается

3. $H_0: M_{\hat{x}} = M_{\hat{y}}; \quad H_1: M_{\hat{x}} < M_{\hat{y}}.$

При этом строят левостороннюю критическую область таким образом, чтобы выполнялось условие

$$p(\hat{U} < U_\alpha) = \alpha.$$

Границу критической области находим из равенства

$$\int_{-\infty}^{U_\alpha} \varphi_{\hat{t}_{(n+m-2)}}(t)dt = \alpha \qquad (4.53)$$

по специальным таблицам Стьюдента.

Величину U определяют по формуле (4.50). Если $U > U_\alpha$, то нулевая гипотеза принимается, в противном случае отвергается.

4.6.2 Проверка гипотез о равенстве дисперсий

Проверка гипотез о равенстве дисперсий – одна из важных задач статистической обработки информации. На практике задача сравнения дисперсий возникает, если требуется сравнить точность приборов, погрешность показаний измерительных устройств, точность методов измерений и т. д.

Пусть имеются две случайные величины $\hat{x}$ и $\hat{y}$, каждая из которых подчиняется нормальному закону распределения с дисперсиями $D_{\hat{x}}$ и $D_{\hat{y}}$. по независимым выборкам ($\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$) и ($\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m$), объем которых соответственно равны n и m, найдены оценки дисперсий:

$$\widehat{\widetilde{D}}_{\hat{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{x}_i - \widehat{\widetilde{m}}_{\hat{x}})^2 = \frac{D_{\hat{x}}}{n-1}\hat{\chi}^2_{<n-1>};$$

$$\widehat{\widetilde{D}}_{\hat{y}} = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{y}_j - \widehat{\widetilde{m}}_{\hat{y}})^2 = \frac{D_{\hat{y}}}{m-1}\hat{\chi}^2_{<m-1>}. \quad (4.54)$$

Обычно полученные оценки различны, в связи с чем возникает вопрос, существенно или несущественно они различаются, т. е. можно ли на основании обработки результатов наблюдений полагать, что $D_{\hat{x}} = D_{\hat{y}}$. (нулевая гипотеза). Если нулевая гипотеза справедлива, то это означает, что выборочные дисперсии (4.54) представляют собой оценки одной и той же числовой характеристики рассеивания генеральной совокупности и их различие определяется случайными причинами. В противном случае (если H_0 отвергнута) различие оценок существенно и является следствием того, что дисперсии генеральных совокупностей различны.

В качестве показателей согласованности о равенстве дисперсий причем отношение оценки дисперсии к меньшей, т. е. величину

$$U = \frac{\widetilde{D}_Б}{\widetilde{D}_M} \quad (4.55)$$

Для определенности положим, что $D_{\hat{x}} > D_{\hat{y}}$, поскольку за счет изменения символов задачи всегда можно обеспечить выполнение данного неравенства. Учитывая оценки (4.54) при условии, что нулевая гипотеза справедлива, на основе отношения (4.55) получаем следующее выражение для показателя согласованности:

$$U = \frac{\hat{\chi}^2_{(n-1)}(m-1)}{\hat{\chi}^2_{(m-1)}(n-1)} = \hat{f}_{(n-1,m-1)}. \quad (4.56)$$

Таким образом, показатель согласованности представляет собой случайную величину, подчиненную закону распределения Фишера со степенями свободы k_1= n – 1, k_2 = m – 1. Напомним, что n – объем выборки, по которой вычислена большая оценка дисперсии, m - объем выборки, по которой вычислена меньшая оценка дисперсии. Как известно, распределение Фишера зависит только от значений степеней свободы и не зависит от других параметров.

Критическая область в зависимости от вида конкурирующей гипотезы:

$$D_{\hat{x}} \neq D_{\hat{y}};\ D_{\hat{x}} > D_{\hat{y}};\ D_{\hat{x}} < D_{\hat{y}}.$$

Построение критических областей для каждого из этих видов осуществляется следующим образом:

1. $H_0: M_{\hat{x}} = M_{\hat{y}}$; $H_1: M_{\hat{x}} \neq M_{\hat{y}}$.

В этом случае строят двустороннюю критическую область, исходя из того, чтобы вероятность попадания в нее показателя согласованности при предположении о справедливости нулевой гипотезы была равна принятому уровню значимости α. При этом наибольшая мощность критерия проверки достигается тогда, когда вероятности попадания показателя согласованности в каждый из двух интервалов критической области будут одинаковы и равны $\alpha/2$. Таким образом, при построении критической области должны выполняться следующие условия (рис. 4.13):

$$\left.\begin{aligned} p(\hat{U} < U_{a_1}) &= \frac{\alpha}{2}, \\ p(U \geq U_{a_2}) &= \frac{\alpha}{2}. \end{aligned}\right\}$$

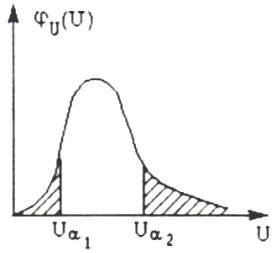


Рис. 4.13

Правая критическая точка U_{a_2} может быть найдена непосредственно по таблице распределения Фишера табл. 4 (приложение). При этом входами в таблицу будут величины $Q = \alpha/2$, $k_1 = n - 1$, $k_2 = m - 1$, т. е.

$$U_{a_2} = f_{(\frac{\alpha}{2}, n-1, m-1)} = f_{a_2}.$$

Однако левых критических точек эта таблица не содержит и поэтому непосредственно по ней найти U_{a_1} невозможно. Это объясняется тем. Что таблицы распределения Фишера составлены лишь для выбора вероятностей из областей малых значений а (практически они составлены для α = 0,01; 0,02; 0,1; 0,2. В связи с этим для нахождения левой критической границы U_{a_1} приходится прибегать к следующему приему. Рассмотрим события $f_{(n-1,m-1)} < f_{\alpha_1}$ и $1/f_{(n-1,m-1)} \geq 1/f_{\alpha_1}$. Так как эти события эквиваленты, то их вероятности равны, т. е.

$$\frac{\alpha}{2} = p(\hat{f}_{(n-1,m-1)} < f_\alpha) = p\left(\frac{1}{\hat{f}_{(n-1,m-1)}} \geq \frac{1}{f_{\alpha_2}}\right).$$

Однако случайная величина $1/\hat{f}_{(n-1,m-1)}$ также подчиняется распределению Фишера со степенями свободы k_1= m – 1, k_2 = n – 1. Поэтому значение $1/f_{\alpha_1}$ может быть найдено как верхний (100 – α/2)%-ный предел этого закона распределения, т. е. $1/f_{\alpha_1} = f_{\frac{\alpha}{2},n-1,m-1}$. Таким образом, для определения $1/f_{\alpha_1}$ необходимо войти в таблицу распределения Фишера с аргументами $Q = \alpha/2$; k_1= m – 1, k_2 = n – 1.

Значение левой критической границы определяется как величина, обратная значению, найденному по таблице.

Учитывая сказанное выше, правило проверки гипотезы о равенстве дисперсий можно сформулировать следующим образом:

а) назначается уровень значимости α и по таблице распределения Фишера находятся критические границы U_{a_1} и U_{a_2}. При нахождении критической границы U_{a_1} входим в таблицу с аргументами $Q = \alpha/2$; k_1= n – 1, k_2 =m– 1, а при определении критической границы U_{a_2} - аргументами $Q = \alpha/2$; k_1= m – 1, k_2 = n – 1. В этом случае табличное значение f_T используется для определения критической границы U_{a_1} из отношения

$$U_{a_1} = \frac{1}{f_\text{T}}; \qquad (4.57)$$

б) вычисляется значение показателя согласованности

$$\text{U} = \frac{\tilde{D}_{\hat{x}}}{\tilde{D}_{\hat{y}}} = \frac{\tilde{\delta}^2_{\hat{x}}}{\tilde{\delta}^2_{\hat{y}}}; \qquad (4.58)$$

в) проверяется неравенство $U_{a_1} < \text{U} < U_{a_2}$. если оно выполняется, т. е. наблюдаемое значение показателя согласованности попадает в область допустимых значений, то делается вывод об отсутствии существенного различия между сравнимыми дисперсиями, т. е. H_0 принимается. если $\text{U} < U_{a_1}$ или $\text{U} > U_{a_2}$, то нулевая гипотеза отвергается.

2. H_0: $D_{\hat{x}} = D_{\hat{y}}$; H_1: $D_{\hat{x}} > D_{\hat{y}}$.

в этом случае строят правостороннюю критическую область таким образом, чтобы вероятность показателя $\widehat{U}$ в эту область при предположении о справедливости нулевой гипотезы была равна принятому уровню значимости:

$$\text{p}(\widehat{U}_\alpha > \text{U}) = \alpha.$$

Критическую точку $U_\alpha = f_{(a,k_1,k_2)}$ находят по таблице распределения Фи-

шера, используя в качестве входов значения $Q = \alpha/2$; k_1= n – 1, k_2 =m– 1. наблюдаемое значение $\widehat{U}$ определяется по формуле (4.58). Если U $<$ U_α, то нет оснований отвергнуть нулевую гипотезу, в противном случае она отвергается.

3. H_0: $D_{\hat{x}} = D_{\hat{y}}$; H_1: $D_{\hat{x}} < D_{\hat{y}}$.

В этом случае строят левостороннюю критическую область таким образом, чтобы

$$p(\widehat{U}_\alpha < U) = \alpha.$$

Критическую точку находят по таблице распределения Фишера, используя отношение

$$U_\alpha = \frac{1}{f_T};$$

где f_T - табличное значение, найденное при входах в таблицу $Q = \alpha/2$; k_1= m – 1, k_2 = n – 1. Наблюдаемое значение $\widehat{U}$ определяется по формуле (4.58). Если U $>$ U_α, то нулевая гипотеза принимается, в противном случае отвергается. Показатель согласованности (4.55) можно использовать для сравнения дисперсий и в том случае, когда одна из них известна, т. е. для нее найдена не оценка, а точное значение. При этом число степеней свободы законы распределения Фишера в числителе или знаменателе выражения (4.56) следует устремить к бесконечности. Методика проверки гипотезы остается прежней, только из степеней свободы (k или k_2) будет иметь бесконечное значение.

5 СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

5.1. Постановка задачи статистического оценивания

Реализация вероятностных моделей каких-либо объектов связана с необходимостью приближенного определения по результатам наблюдений – (статистического оценивания) – соответствующих характеристик рассматриваемых случайных объектов (событий, величин, векторов, функций). Задачи подобного рода особенно часто встречаются в метеорологической практике. Так, при прогнозировании гроз могут быть использованы результаты статистического оценивания вероятности грозы при данном исходном состоянии атмосферы, при регрессионном прогнозе температуры воздуха – результаты статистического оценивания вектора регрессии и т. д.

В теории статистического оценивания предполагается, что «правило» оценивания рассматриваемой характеристики может быть использовано для любых случайных (репрезентативных) выборок из генеральной совокупности. Поэтому оценки этой характеристики, полученные в результате применения указанного правила к конкретным выборкам, считаются реализациями соответствующей случайной величины (вектора, функции), близость которых к оцениваемой характеристике должна пониматься в вероятностном смысле.

Знакомство с теорией статистического оценивания начнем с обсуждения вопросов, связанных с оцениванием числовых характеристик распределения случайных величин. Как известно, исчерпывающим описанием случайной величины является закон ее распределения. Однако для решения многих прикладных задач бывает достаточным знание только нескольких числовых характеристик этого распределения, приближенные значения которых и должны быть определены в результате статистического оценивания.

Обозначим через a оцениваемую числовую характеристику распределения случайной величины $\hat{x}$, а через $x_1, x_2, \ldots, x_n$ - значения $\hat{x}$, полученные в n испытаниях. В силу случайности выборки эти значения могут рассматриваться как реализации независимых и одинаково распределенных с $\hat{x}$ случайных величин $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$.

Пусть $\hat{a} = f(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ – некоторая функция случайной выборки (статистика), принимающая для конкретных выборок значения $a_1^*, a_2^*, \ldots$. Основная задача оценивания заключается в определении такой статистики a^*, которая обеспечивала бы близость «в среднем» оценок $a_1^*, a_2^*, \ldots, ka$.

Очевидно, что решению указанной задачи должно предшествовать уточнению понятия близость «в среднем» оценок к оцениваемой характеристике и за-

дание требований к оценкам a^* , руководствуясь которыми можно было бы выбирать оценки, лучше в указанном смысле.

5.2. Требование к статистическим оценкам

В качестве меры близости оценок $\hat{a}^*$ к a в теории статистического описания принято использовать математическое ожидание квадрата разности $\hat{a}^*$ - a (близость « в среднеквадратическом»). Лучшими признаются оценки соответствующие меньшим значениям $\mathrm{M}[(\hat{a}^* - a)^2]$.

Поскольку

$$\mathrm{M}[(\hat{a}^* - a)^2] = \mathrm{M}\{[(\hat{a}^* - \mathrm{M}_{\hat{a}^*})^2]\} =$$

$$= \mathrm{M}[(\hat{a}^* - \mathrm{M}_{\hat{a}^*})^2] + 2(\hat{a}^* - \mathrm{M}_{\hat{a}^*})(\mathrm{M}_{\hat{a}^*} - a) + (\mathrm{M}_{\hat{a}^*} - a)^2, \quad (5.1)$$

$$\mathrm{M}\,[2(\hat{a}^* - \mathrm{M}_{\hat{a}^*})(\mathrm{M}_{\hat{a}^*} - a)] = 0 \text{ и } \mathrm{M}\,[(\hat{a}^* - \mathrm{M}_{\hat{a}^*})]^2 = D_{\hat{a}^*},$$

$$\mathrm{M}[(\hat{a}^* - a)^2] = D_{\hat{a}^*} + (\mathrm{M}_{\hat{a}^*} - a)^2.$$

Разность $(\mathrm{M}_{\hat{a}^*} - a)$ характеризует смещение оценок, «систематическую ошибку» оценивания. Из множества оценок (статистик) с одинаковыми дисперсиями $D_{\hat{a}^*}$ лучшей должна быть признана оценка с минимальным смещением $(\mathrm{M}_{\hat{a}^*} - a)$.

Оценки, для которых $(\mathrm{M}_{\hat{a}^*} = a)$ и, следовательно,

$$(\mathrm{M}_{\hat{a}^*} - a)^2 = 0, \quad (5.2)$$

называются несмещенными.

С другой стороны, среди оценок с одинаковым смещением лучшей должна быть признана оценка (статистика) с минимальной дисперсией $D_{\hat{a}^*}$. Несмещенные оценки, обеспечивающие выполнение условия

$$D_{\hat{a}_k^*} = \min_{\hat{a}^*} D_{\hat{a}^*}, \quad (5.3)$$

называются эффективными**. ***

Требования несмещенности и эффективности относятся к оценкам полученным по выборкам фиксированного объема n. Следующие два требования характеризуют поведение оценок при увеличении объема выборок. Очевидно, что с увеличением объема выборок возрастает количество содержащейся в них информации о распределении $\hat{x}$ и, в частности, о значении оцениваемой характеристики a. Логично поэтому потребовать, чтобы с увеличением n оценки $a^*_{(n)}$ «приближались» к a.

---

* Следует отметить, что иногда за счет небольшого смещения оценок удается добиться значительного уменьшения дисперсий, и такие оценки оказываются в смысле формулы (5.1) предпочтительнее эффективных, определенных в классе несмещенных оценок.

Оценки, сходящиеся по вероятности к оцениваемой числовой характеристике, называются состоятельными. Требование состоятельности оценки $a^*_{(n)}$ может быть записано в виде равенства

$$\lim_{n\to\infty} P(\left|a^*_{(n)} - a\right| \le \varepsilon = 1, \tag{5.4}$$

где n – объем выборки,

$P(\left|a^*_{(n)} - a\right| \le \varepsilon$ – вероятность события, указанного в скобках,

ε – произвольное положительное число.

Оценки $\hat{a}^*$, удовлетворяющие всем трем перечисленным требованиям называются подходящими значениями a.

Как показывает практика статистического оценивания, наиболее трудно выполнимым обычно является требование эффективности (4.3), поэтому часто подходящими значениями a считаются оценки, для которых условие (4.3) выполняется только в пределе, при $n \to \infty$. Такие оценки называют асимптотически эффективными.

5.3. Методы статистического оценивания

В основу методов определения статистик $\hat{a}^*$, удовлетворяющих сформулированным ранее требованиям, могут быть положены различные принципы, наиболее распространенными среди которых являются принципы равенства моментов распределения (РМР) и принципы максимума правдоподобия (МП). Методы, основанные на этих двух принципах, и будут рассмотрены далее. Некоторые сведения о других принципах и методах оценивания будут даны в подразд. 6.2.

5.3.1.Метод моментов

Пусть оцениваемой числовой характеристикой распределения случайной величины $\hat{x}$ является начальный момент порядка k

$$v_k[\hat{x}] = M[\hat{x}^{\,k}].$$

Будем считать, что элементы случайной выборки $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ - независимые случайные величины, распределенные одинаково с $\hat{x}$. Тогда случайные величины $\hat{x}_1^k, \hat{x}_2^k, \ldots, \hat{x}_n^k$ независимы и имеют то же распределение, что $\hat{x}^{\,k}$. Поэтому по теореме Чебышева

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i^k - M_{\hat{x}^k}\right| < \varepsilon\right) = 1$$

и следовательно, выборочный (статистический) момент порядка k

$$\hat{v}_k^*\,[\hat{x}] = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i^k \tag{5.5}$$

находится по вероятности к $v_k[\hat{x}]$.

Кроме того, из формулы 5.5 следует, что

$$M\{\hat{v}_k^*\,[\hat{x}]\} = M[\hat{x}^{\,k}] = v_k[\hat{x}].$$

Таким образом, выборочные моменты (4.5) являются состоятельными несмещенными оценками соответствующих моментов $\hat{x}$.

Метод оценивания, основанный на предположении, что

$$\hat{v}_k^* [\hat{x}] = v_k[\hat{x}],$$

принято называть методом равных моментов или просто методом моментов, а полученные этим методом оценки – РМ – оценками.

Второе принципиальное предположение, используемое в методе моментов, заключается в следующем. Представим оцениваемую числовую характеристику a в виде функции от начальных моментов: $a = f(v_1, v_2, \ldots, v_m)$, будем считать РМ – оценкой a ту же функцию f от соответствующих выборочных моментов:

$$a^* = f(v_1^*, v_2^*, \ldots, v_m^*). \qquad (5.6)$$

Оценки (5.6) являются состоятельными, но не обязательно несмещенными и эффективными.

Пример 1. Пользуясь методом моментов, оценить математическое ожидание и дисперсию случайной величины $\hat{x}$.

По определению $M_{\hat{x}} = v_1[\hat{x}]$, $D_{\hat{x}} = M[(\hat{x} - M_{\hat{x}})^2] = M_{\hat{x}^2} - M_{\hat{x}}^2 =$

$= v_2[\hat{x}] - v_1^2[\hat{x}]$.

Следовательно РМ – оценки $M_{\hat{x}}$ и $D_{\hat{x}}$ имеют вид:

$$M_{\hat{x}}^* = v_1^*[\hat{x}] = \frac{1}{n}\sum_{i=1}^n x_i \qquad (5.7)$$

$$D_{\hat{x}}^* = v_2^*[\hat{x}] - v_1^{*2}[x] = \frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(x_i - \frac{1}{n}\sum_{i=1}^n x_i\right)^2. \qquad (5.8)$$

Как будет показано в подразд. 5.6. и 5.7, оценка (5.7) является несмещенной, а оценка (5.8) – смещена, причем для устранения смещения достаточно умножить (5.8) на отношение n/n – 1, то есть несмещенной является оценка

$$D_{\hat{x}}^* = \frac{n}{n-1}\sum_{i=1}^n \left(x_i - \frac{1}{n}\sum_{i=1}^n x_i\right)^2. \qquad (5.9)$$

5.3.2 Метод максимального правдоподобия

Для реализации метода моментов не требуется какой-либо априорной информации о распределении случайной величины. В методе максимального правдоподобия оцениваемыми характеристиками являются параметры, законов распределения, известных aprioi « с точностью до оцениваемых параметров». В основе метода лежат следующие соображения. Пусть плотность распределения случайной величины $\hat{x}$ в точке x есть известная функция от x и оцениваемого

параметра a: $\varphi_{\hat{x}}(x, a)$. Будем рассматривать результаты независимых наблюдений $x_1, x_2, \ldots, x_n$, как компоненты вектора $X_{<n>}$. Тогда плотность распределения случайного вектора $\widehat{X}_{<n>}$

$$\varphi\widehat{X}_{<n>}(X_{<n>}; a) = \prod_{i=1}^{n} \varphi_{\hat{x}}(x_i; a) \qquad (5.10)$$

при заданном $X_{<n>}$ определяется только значением параметра a. В методе максимального правдоподобия в качестве статистической оценки a^*

используется то значение, при котором функция (5.10) максимальна, то есть получение выборки $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ «более вероятно», чем при любых других значениях a. *)

Оценка a^* параметра a, доставляющие максимум функции правдоподобия (5.10), называются оценками максимального правдоподобия или МП-оценками. Для нахождения МП-оценки параметра a необходимо решить относительно a уравнение правдоподобия

$$\frac{\partial}{\partial a}\prod_{i=1}^{n} \varphi_{\hat{x}}(x_i; a) = 0. \qquad (5.11)$$

Так как минимумы функций $\varphi_{\widehat{X}_{<n>}}(X_{<n>}; a)$ и $\ln \varphi_{\widehat{X}_{<n>}}(X_{<n>}; a)$ достигаются при одном и том же значении a, вместо уравнения (5.11) можно для определения a^* использовать уравнение

$$\frac{\partial}{\partial a}\sum_{i=1}^{n} \ln \varphi_{\hat{x}}(x_i; a) = 0. \qquad (5.12)$$

Если оцениванию подлежит не один, а несколько параметров, необходимо решение системы уравнений правдоподобия, полученных приравниванием нулю частных производных функции правдоподобия по каждому из оцениваемых параметров. МП-оценки, как правило. Состоятельны, асимптотически эффективны и, особенно для выборок малого объема, более предпочтительны, чем соответствующие РМ-оценки.

Пример 2. Пользуясь методом максимального правдоподобия, оценить параметры $M_{\hat{x}}$ и $\delta_{\hat{x}}$ нормального закона распределения случайной величины $\hat{x}$. плотность Z нормального распределения $\hat{x}$ определяется формулой

$$\varphi_{\hat{x}}(X, M_{\hat{x}}, \delta_{\hat{x}}) = \frac{1}{\delta_{\hat{x}}\sqrt{2\pi}}\, e^{\frac{-(X-M_{\hat{x}})^2}{2\delta_{\hat{x}}^2}} \qquad (5.13)$$

Уравнение правдоподобия получаем подстановкой выражения для $\varphi_{\hat{x}}$ в формулу (5.12):

$$\frac{\partial}{\partial M_{\hat{x}}}\sum_{i=1}^{n} \ln \varphi_{\hat{x}}(x; M_{\hat{x}}; \delta_{\hat{x}}) = \\ = \frac{\partial}{\partial M_{\hat{x}}}\sum_{i=1}^{n}\left[-ln\ \delta_{\hat{x}} - \ ln\sqrt{2\pi} - \frac{(x-M_{\hat{x}})^2}{2\delta_{\hat{x}}^2}\right] = 0 \qquad (5.14)$$

$$\frac{\partial}{\partial \delta_{\hat{x}}}\sum_{i=1}^{n} \ln \varphi_{\hat{x}}(x; M_{\hat{x}}; \delta_{\hat{x}}) = \\ = \frac{\partial}{\partial \delta_{\hat{x}}}\sum_{i=1}^{n}\left[-ln\ \delta_{\hat{x}} - \ ln\sqrt{2\pi} - \frac{(x-M_{\hat{x}})^2}{2\delta_{\hat{x}}^2}\right] = 0 \qquad (5.15)$$

Из (5.14) следует $\sum_{i=1}^{n} \frac{x_i - M_{\hat{x}}}{\delta_{\hat{x}}^2} = 0$ и $M_{\hat{x}}^* = \frac{1}{n}\sum_{i=1}^{n} x_i$, а из (5.15) $\sum_{i=1}^{n}\left[-\frac{1}{\delta_{\hat{x}}} + \right.$
$\left. + \frac{(x - M_{\hat{x}})^2}{\delta_{\hat{x}}^2}\right] = 0$ и $\delta_{\hat{x}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - M_{\hat{x}}^*)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)}$.

5.4. Анализ качества оценивания

Ранее рассматривалась задача определения статистик $\hat{a}$, значения которых (оценки) могли бы считаться приблизительно равными значениями оцениваемой характеристики a. Такие оценки принято называть точечными. «Качество» точечных оценок $\hat{a}^*$ естественно характеризовать их близость к a, причем в силу случайности статистик $\hat{a}^*$, эта близость должна пониматься в вероятностном смысле. Вопросы анализа качества точечных оценок, близости к a составляют основное содержание теории интервального оценивания.

Введем некоторые понятия, используемые в теории интервального оценивания. Вероятность β того, что оцениваемый параметр имеет значение, находящееся в интервале $[\hat{a}_1 \le a \le \hat{a}_2]$,

$$\beta = P(\hat{a}_1 \le a \le \hat{a}_2), \tag{5.16}$$

называется доверительной вероятностью, интервал $[\hat{a}_1, \hat{a}_2]$ – доверительным интервалом, а его границы $\hat{a}_1$ и $\hat{a}_2$ - доверительными границами. В геометрической интерпретации доверительная вероятность – вероятность «накрытия» точки a отрезком с концами в случайных точках $\hat{a}_1$ и $\hat{a}_2$. При анализе качества точечного оценивания доверительные границы представляются в виде:

$$\begin{aligned} \hat{a}_1 &= \hat{a}_2^* - \varepsilon_1, \\ \hat{a}_2 &= \hat{a}^* + \varepsilon_2. \end{aligned} \tag{5.17}$$

Тогда

$$\beta = P(\hat{a}^* - \varepsilon_1, \le a \le \hat{a}^* + \varepsilon_2), \tag{5.18}$$

Для несмещенных оценок симметрично распределенных относительно a, обычно принимают $\varepsilon_1 = \varepsilon_2 = \varepsilon$ и

$$\beta = P(|a - \hat{a}^*| \le \varepsilon), \tag{5.19}$$

где ε – полуширина доверительного интервала $[\hat{a}^* - \varepsilon, \hat{a}^* + \varepsilon]$.

Если величину $|a - \hat{a}^*|$ считать мерой близости оценок к a, то соотношение (5.19) может быть использовано для характеристики точности оценивания – по значению ε при заданной доверительной вероятности β и надежности (достоверности) оценивания – по значению βпри заданной полуширине доверительного интервала ε. Точность и надежность оценивания, как правило, повышаются с увеличением объема выборок. Минимальный объем выборки, обеспечивающий оценивание рассматриваемой характеристики с заданными точностью и надежностью, называется потребным объемом.

Для анализа точности и надежности оценивания, очевидно, необходимо

знание закона распределения $\hat{a}^*$. Рассмотрим ситуацию, когда при больших объемах выборок n распределение a близко к нормальному с параметрами $M_{\hat{a}^*}$ и $\delta_{\hat{a}^*}$. Введем новую случайную величину $\hat{z} = \frac{\hat{a}^* - a}{\delta_{\hat{a}^*}}$.

В силу несмещенности оценок $\hat{a}^* M_{\hat{x}} = 0$. $M_{\hat{z}} = 0$ и $\delta_Z = 1$, причем распределение случайной величины z (линейно зависящей от $\hat{a}^*$) тоже нормально.

$$\beta = P(|a - \hat{a}^*| \le \varepsilon) = P(\delta_{\hat{a}^*} |\hat{z}| \le \varepsilon) = P(|\hat{z}| \le \frac{\varepsilon}{\delta_{\hat{a}^*}}) \approx$$
$$\approx F_{\hat{z}}^{[H]}\left(\frac{\varepsilon}{\delta_{\hat{a}^*}}\right) - F_{\hat{z}}^{[H]}\left(\frac{-\varepsilon}{\delta_{\hat{a}^*}}\right) = 2\varphi_0\left(\frac{\varepsilon}{\delta_{\hat{a}^*}}\right), \quad (5.20)$$

где $F_{\hat{z}}^{[H]}(z)$ – функция распределения для нормального закона

$\varphi_0(z)$ – интеграл вероятностей.

Таким образом, если для оценивания числовой характеристики a используется статистика $\hat{a}^*$, удовлетворяющая требованию несмещенности и распределенная асимптотически нормально, то при больших n выполняется

$$\beta = 2\varphi_0\left(\frac{\varepsilon}{\delta_{\hat{a}^*}}\right). \quad (5.21)$$

В практике оценивания величина $\delta_{\hat{a}^*}$, как правило, неизвестна и заменяется в (5.21) ее статистической оценкой $\delta_{\hat{a}^*}$, в результате чего формула принимает вид:

$$\beta \approx 2\varphi_0\left(\frac{\varepsilon}{\delta_{\hat{a}^*}}\right). \quad (5.22)$$

Значения интеграла вероятности табулированы и приводятся в приложении (табл. 1).

5.5. Оценивание вероятности случайного события

Пусть случайное событие $\hat{A}$ в n независимых испытаний осуществилось m раз. Назовем отношение $\frac{m}{n}$ частотой появления события в n испытаниях или статистической вероятностью этого события.

Рассмотрим возможность использования для оценивания вероятности $P(\hat{A}) = p$ статистики

$$\hat{p}^* = \frac{m}{n} \quad (5.23)$$

Начнем с определения математического ожидания и дисперсии $\hat{p}^*$. Обозначим $m_{(i)}$ число осуществлений события А в одном i-м испытании. Так как случайная величина $\hat{m}_{(i)}$ принимает значение 1 с вероятностью p и значение 0 с вероятностью (1-p), тогда

$$M_{\hat{m}} = p\,1 + (1-p)0 = p, \quad (5.24)$$
$$D_{\hat{m}_{(i)}} = p(1-p)^2 + (1-p)(0-p)^2 = p(1-p). \quad (5.25)$$

Но $\hat{m} = \sum_{i=1}^{n} \hat{m}_{(i)}$, поэтому

$$M_{\hat{p}^*} = \frac{1}{n} M\left[\sum_{i=1}^{n} \hat{m}_{(i)}\right] = p, \quad (5.26)$$

$$D_{\hat{p}^*} = \frac{1}{n^2} D\left[\sum_{i=1}^{n} \hat{m}_{(i)}\right] = \frac{p(1-p)}{n}. \quad (5.27)$$

Посмотрим теперь, насколько при использовании для оценивания вероятности p статистики $\hat{p}^* = \frac{\hat{m}}{n}$ выполняются требования, сформулированные в подразд. 5.2.

Как следует из (4.26), $\frac{m}{n}$ - несмещенная оценка p. Кроме того, эта оценка и состоятельна, так как по теореме Бернули

$$\lim_{n\to\infty} P\left(\left|\frac{\hat{m}}{n} - p\right| < \varepsilon\right) = 1. \quad (5.28)$$

Наконец, рассматриваемая оценка асимптотически эффективна, поскольку

$$\lim_{n\to\infty} D_{\hat{p}^*} = \lim_{n\to\infty} \frac{p(1-p)}{n} = 0. \quad (5.29)$$

Более того, можно показать, что статистика $\frac{\hat{m}}{n}$ обладает минимальной дисперсией при любом n. Таким образом, частота $\hat{p}^* = \frac{\hat{m}}{n}$ - подходящее значение вероятности p.

Перейдем к анализу качества оценивания. По теореме Муавра-Лапласа биномиальный закон, по которому распределено $\hat{m}$, с увеличением n приближается к нормальному закону с параметрами $M_{\hat{m}}$ и $\delta_{\hat{m}}$. Так как $\hat{p}^*$ - линейная функция $\hat{m}$, распределение $\hat{p}^*$ при больших n тоже должно быть приближено нормальным. Следовательно, при анализе качества оценивания p по выборкам большого объема может быть использовано соотношение (5.22), которое с учетом (5.27) принимает вид:

$$\beta \approx 2\phi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \approx 2\phi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p^*(1-p^*)}}\right). \quad (5.30)$$

Полученная формула может быть использована для решения трех основных задач анализа качества оценивания p:

- определения доверительной вероятности β при заданных объеме выборки n и ширине доверительного интервала 2ε;

- определения ширины доверительного интервала 2ε при заданных объеме выборки n и доверительной вероятности β;

- определение потребного объема выборки при заданных ширине доверительного интервала 2ε и доверительной вероятности β.

Пример 3. Определить минимальный объем выборки, достаточный для оценивания вероятности случайного события p с заданными точностью ($\varepsilon = 0{,}01$) и надежностью ($\beta = 0{,}95$), если приближенное значение p равно 0,4.

Подставим заданные значения β, ε и p в формулу (5.30): $0{,}95 = 2\phi_0\left(\frac{0{,}01\sqrt{n}}{0.4\ 0.6}\right)$.

Для $\phi_0(x) = 0{,}475$ по табл. 1 Приложения найдем $x = \left(\frac{0{,}01\sqrt{n}}{0.4\ 0.6}\right) = 1{,}96$, отсюда

потребный объем n = 9220.

Отметим, что потребный объем существенно снижается при удалении p от 0,5, увеличении ε и уменьшении β. Так, при p = 0,1 (или 0,9), ε = 0,05 и той же доверительной вероятности β потребный объем равен 139, а при β = 0,90 – всего 98.

Пример 4. В табл. 5.1 приводятся средние значения температуры воздуха T, в Санкт-Петербурге в марте 1881 г., (i=1), 1882 г. (i=2), …, 1935 г. (i=55). Оценить вероятность p того, что средняя мартовская температура положительна ($\hat{\mathrm{T}} > 0$°C).

Таблица 5.1

i	1	2	3	4	5	6	7	8
T_i(°C)	-6.4	-0.1	-8.2	-5.4	-3.9	-5.9	-3.9	-10.1
i	9	10	11	12	13	14	15	16
T_i(°C)	-7.3	0.4	-2.8	-5.0	-5.0	-3.0	-5.0	-3.0
i	17	18	19	20	21	22	23	24
T_i(°C)	-4.5	-6.8	-8.1	-5.6	-5.7	-5.2	0.4	-3.9
i	25	26	27	28	29	30	31	32
T_i(°C)	-1.3	-4.3	-2.6	-4.7	-3.3	0.2	-3.6	0.6
i	33	34	35	36	37	38	39	40
T_i(°C)	-0.8	-2.6	-8.5	-4.5	-11.1	-4.7	-7.4	0.8
i	41	42	43	44	45	46	47	48
T_i(°C)	0.6	-3.9	-5.3	-4.6	-4.6	-4.2	-1.8	-4.4
i	49	50	51	52	53	54	55	
T_i(°C)	-5.0	-1.5	-6.5	-6.5	-3.0	-2.0	-2.9	

Используем в качестве оценки p повторяемость p^* вариантов с $T_i > 0$(°C): $p^* = \frac{m}{n} = \frac{6}{55} = 0{,}109$. Для характеристики качества оценивания найдем доверительную вероятность, соответствующую доверительному интервалу [0,09; 0,209]. По формуле (5.30) при n = 55, p^* = 0,109 и ε = 0,100 получим $\beta \approx 2\Phi_0\left(\frac{0{,}100\sqrt{55}}{\sqrt{0{,}109(1-0{,}109)}}\right) \approx 2\Phi_0(2{,}380)$. По таблицам интеграла вероятности $\Phi_0(2{,}380) = 0{,}491$. Следовательно, β = 2 0,491 = 0,982.

5.6 Оценивание математического ожидания случайной величины

Рассматривается задача статистического оценивания математического ожидания случайной величины $\hat{\mathrm{x}}$. Результаты наблюдений x_1, x_2, …, x_n считаются реализациями независимых случайных величин $\hat{\mathrm{x}}_1$, $\hat{\mathrm{x}}_2$, …, $\hat{\mathrm{x}}_n$ распределенных одинаково с $\hat{\mathrm{x}}$.

В примере 2 подразд. 5.3.2 получена МП-оценка $\mathrm{M}_{\hat{\mathrm{x}}}$:

$$\mathrm{M}^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (5.31)$$

Нетрудно показать, что статистика

$$\widehat{M}^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (5.32)$$

удовлетворяет основным требованиям, сформулированным в подраз. 5.2. Действительно:

$$M[\widehat{M}^*_{\hat{x}}] = \frac{1}{n} M[\sum_{i=1}^{n} x_i] = M_{\hat{x}}. \qquad (5.33)$$

$$\lim_{n\to\infty} D[\widehat{M}^*_{\hat{x}}] = \lim_{n\to\infty} \frac{1}{n^2} D[\sum_{i=1}^{n} x_i] = \lim_{n\to\infty} \frac{D_{\hat{x}}}{n} = 0. \qquad (5.34)$$

следовательно, оценки (5.31) являются несмещенными и асимптотически эффективными. Для некоторых распределений $\hat{x}$, и в частности для нормального распределения, оценки (5.31) обеспечивают минимальную дисперсию при любом n, т. е. эффективны.

Наконец, по теореме Чебышева

$$\lim_{n\to\infty} p\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - M_{\hat{x}}\right| < \varepsilon\right) = 1, \qquad (5.35)$$

и оценки (5.31) состоятельны. Таким образом, оценка $\widehat{M}^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (выборочное среднее или статистическое математическое ожидание $\hat{x}$) – подходящее значение $M_{\hat{x}}$. Для анализа качества оценивания учтем, что в силу центральной предельной теоремы распределение $\widehat{M}^*_{\hat{x}}$ при больших n близко к нормальному с параметрами $M[\widehat{M}^*_{\hat{x}}] = M_{\hat{x}}$ и в $[\widehat{M}^*_{\hat{x}}] = \frac{\delta_{\hat{x}}}{\sqrt{n}}$. Поэтому, воспользовавшись формулой (5.29), получим:

$$\beta \approx 2\phi_0\left(\frac{\varepsilon\sqrt{n}}{\delta_{\hat{x}}}\right) \approx 2\phi_0\left(\frac{\varepsilon\sqrt{n}}{\delta^*_{\hat{x}}}\right). \qquad (5.36)$$

Пример 5. Требуется по результатам наблюдений, представленным в табл. 5.1, оценить $M_{\hat{T}}$.

По формуле (5.32) $M^*_{\hat{T}} = \frac{1}{55}\sum_{i=1}^{55} T_i$ = -4,13°C. Среднее квадратическое отклонение $\hat{T}$, как отмечалось в п. 5.3.1, оценивается величиной

$$\delta^*_{\hat{T}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(T_i - \frac{1}{n}\sum_{i=1}^{n} T_i)^2} = \sqrt{\frac{1}{54}\sum_{i=1}^{55}(T_i - 4{,}13)^2} = 2{,}69°C. \qquad (5.37)$$

Поэтому

$$\beta \approx 2\phi_0\left(\frac{\varepsilon\sqrt{55}}{2{,}69}\right) \approx 2\phi_0(2{,}757\,\varepsilon). \qquad (5.38)$$

Формулу (5.37) можно использовать для анализа точности и надежности оценивания. Определим, например. Границы доверительного интервала при β = 0,95. По таблице интеграла вероятностей найдем 2,757 ε = 1,96, следовательно, ε = 0,711°C, и с вероятностью 0,95 $M_{\hat{T}}$ находится в интервале [-4,84; -3,42]°C. Формула (5.30) дает удовлетворительные результаты при близости оценки $\delta^*_{\hat{x}}$ и $\delta_{\hat{x}}$.

Но качество оценивания, как отмечалось, снижается уменьшением объема выборок, поэтому при малых n использование формулы (5.30) нежелательно, и вопросы анализа качества оценивания $M_{\hat{x}}$ в случае малых выборок требуют специального рассмотрения. В настоящее время эти вопросы исследованы только для некоторых «стандартных» распределений $\hat{x}$. Мы ограничимся достаточно часто встречающейся в метеорологии ситуацией, когда распределение $\hat{x}$ подчинено нормальному закону. Как раньше, будем считать элементы случайной выборки $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$, независимыми случайными величинами, распределенными одинаково с $\hat{x}$, но теперь это распределение нормально.

Для оценивания $M_{\hat{x}}$ снова воспользуемся статистикой (5.32):

$$M^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Так как распределение $M^*_{\hat{x}}$ симметрично относительно $M_{\hat{x}}$, примем $\varepsilon_1 = \varepsilon_2 = \varepsilon$. Тогда доверительная вероятность β определяется формулой:

$$\beta = P(|M^*_{\hat{x}} - M_{\hat{x}}| < \varepsilon) = P\left(\frac{|M^*_{\hat{x}} - M_{\hat{x}}|}{\delta^*_{\hat{x}}} < \frac{\varepsilon}{\delta^*_{\hat{x}}}\right). \quad (5.38)$$

В подразд. 5.7 будет показано, что

$$\hat{\delta}^*_{\hat{x}} = K_{n-1}\frac{\delta_{\hat{x}}}{\sqrt{n-1}}\hat{\chi}_{n-1}, \quad (5.39)$$

где $\hat{\chi}_{n-1}$ - случайная величина, имеющая с – распределение с (n – 1) степенями свободы. При малых n можно принять $K_{n-1} \approx 1$. Поэтому

$$\beta = P\left(\frac{|M^*_{\hat{x}} - M_{\hat{x}}|}{\frac{\delta_{\hat{x}}}{\sqrt{n-1}}\chi_{n-1}} < \frac{\varepsilon}{\delta^*_{\hat{x}}}\right) = P\left(\frac{\frac{|M^*_{\hat{x}} - M_{\hat{x}}|}{\delta_{\hat{x}}/\sqrt{n}}}{\frac{\hat{\chi}_{n-1}}{\sqrt{n-1}}} < \frac{\varepsilon\sqrt{n}}{\delta^*_{\hat{x}}}\right) \quad (5.40)$$

По принятому предположению случайные величины $\hat{x}_i$ распределены нормально, следовательно, нормальное распределение должны иметь и линейно связанные с ними величины $M^*_{\hat{x}}$ и

$$\hat{v} = \frac{|M^*_{\hat{x}} - M_{\hat{x}}|}{\delta/\sqrt{n}}, \quad (5.41)$$

причем $M_{\hat{v}} = 0$ и $\delta_{\hat{v}} = 1$.

В курсе теории вероятностей показывается, что случайная величина

$$\frac{\hat{v}}{\chi_{n-1}/\sqrt{n-1}} = \hat{t}_{n-1} \quad (5.42)$$

имеет распределение Стьюдента с (n – 1) степенями свободы, симметричной относительно нуля. Таким образом,

$$\beta = P\left(|\hat{t}_{n-1}| < \frac{\varepsilon\sqrt{n}}{\delta^*_{\hat{x}}}\right) = F^{(c)}_{\hat{t}_{(n-1)}}\left(\frac{\varepsilon\sqrt{n}}{\delta^*_{\hat{x}}}\right) - F^{(c)}_{\hat{t}_{(n-1)}}\left(\frac{-\varepsilon\sqrt{n}}{\delta^*_{\hat{x}}}\right),$$

и, поскольку $F^{(c)}_{\hat{t}_{(n-1)}}\left(\frac{-\varepsilon\sqrt{n}}{\delta^*_{\hat{x}}}\right) = \frac{1}{2} - \frac{\beta}{2}$, окончательно получаем

$$\beta = 2F^{(c)}_{\hat{t}_{(n-1)}}\left(\frac{-\varepsilon\sqrt{n}}{\delta^*_{\hat{x}}}\right) - 1. \qquad (5.43)$$

На основании (5.43) могут быть решены все три основных задачи анализа качества статистического оценивания математического ожидания случайных величин, распределенных по нормальному закону. В справочной литературе чаще публикуются таблицы значений не $F^{(c)}_{\hat{t}_{(k)}}(t_{a;k})$, а обратной функции $t_{a;k} = F^{(c)^1}_{\hat{t}_{(k)}}\left(\frac{1+a}{2}\right)$. (табл. 2 Приложения).

Пример 6. В таблице 5.2 приведены ошибки D_{p_i} десяти прогнозов давления в центре циклонов, составленных с суточной заблаговременностью методом формальной экстраполяции. Требуется оценить $M_{\Delta\hat{p}}$.

Таблица 5.2

i	1	2	3	4	5	6	7	8	9	10
D_{p_i}	5	7	-4	0	10	-6	8	4	3	-1

МП – оценка $M_{\Delta\hat{p}}$ равна

$$M^*_{\Delta\hat{p}} = \frac{1}{10}\sum_{i=1}^{10}\Delta p_i = 2{,}6 \text{ мбар.}$$

а

$$\delta^*_{\Delta\hat{p}} = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(\Delta p_i - M^*_{\Delta\hat{p}})^2} = 5{,}25 \text{ мбар.} \quad *)$$

Следовательно,

$$\beta = 2F^{(c)}_{\hat{t}_{(n-1)}}\left(\frac{-\varepsilon\sqrt{10}}{5{,}25}\right) - 1 \text{ и } 0{,}60\varepsilon = F^{(c)^1}_{t_{(9)}}\left(\frac{1+\beta}{2}\right).$$

Обратившись к табл. 2 Приложения, найдем, например для доверительной вероятности $0{,}9F^{(c)^1}_{t_{(9)}} = 1{,}833$. Таким образом, $\varepsilon = 2{,}29$ мбар, и границами доверительного интервала для $M_{\Delta\hat{p}}$ являются значения 0,31 и 4,89 мбар. Интересно отметить, то при том же значении $\beta = 0{,}9$ полуширина доверительного интервала, рассчитанная по формуле (5.36), оказывается равной 2,74 мбар.

5.7 Оценивание дисперсии и среднего квадратического отклонения случайной величины

Пусть элементы случайной выборки $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ - независимые случайные величины, распределенные так же, как $\hat{x}$. требуется получить оценки характеристик рассеяния $\hat{x}$: $D_{\hat{x}}$ и $\delta_{\hat{x}}$. Рассмотрим наиболее важный для метеорологической практики случай, когда случайная величина $\hat{x}$ распределена по нормальному закону. Начнем с задачи оценивания дисперсии $D_{\hat{x}}$. Как было показано в подразд. 5.3.2, МП – оценкой дисперсии $D_{\hat{x}}$ является статистическая (выборочная) дисперсия

$$D^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathrm{M}^*_{\hat{x}})^2, \qquad (5.44)$$

где $\mathrm{M}^*_{\hat{x}} = \frac{1}{2}\sum_{i=1}^{n} x_i$.

Проверим удовлетворяет ли оценка (5.44) требованию несмещенности. Для этого преобразуем выражение (5.44) следующим образом:

$$D^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}[(\hat{x}_i - \mathrm{M}_{\hat{x}}) - (\mathrm{M}_{\hat{x}} - \widehat{\mathrm{M}}^*_{\hat{x}})]^2 =$$

$$= \frac{1}{n}[\sum_{i=1}^{n}(\hat{x}_i - \mathrm{M}_{\hat{x}})^2 - 2\sum_{i=1}^{n}(\hat{x}_i - \mathrm{M}_{\hat{x}})\left(\mathrm{M}_{\hat{x}} - \widehat{\mathrm{M}}^*_{\hat{x}}\right) + \sum_{i=1}^{n}(\mathrm{M}_{\hat{x}} - \widehat{\mathrm{M}}^*_{\hat{x}})^2\,].$$

Поскольку

$$-2\sum_{i=1}^{n}(\hat{x}_i - \mathrm{M}_{\hat{x}})\left(\mathrm{M}_{\hat{x}} - \widehat{\mathrm{M}}^*_{\hat{x}}\right) = -2\left(\mathrm{M}_{\hat{x}} - \widehat{\mathrm{M}}^*_{\hat{x}}\right)(n\widehat{\mathrm{M}}^*_{\hat{x}} - n\mathrm{M}_{\hat{x}}) =$$

$$= 2\left(\mathrm{M}_{\hat{x}} - \widehat{\mathrm{M}}^*_{\hat{x}}\right)^2.$$

$$\sum_{i=1}^{n}\left(\mathrm{M}_{\hat{x}} - \widehat{\mathrm{M}}^*_{\hat{x}}\right)^2 = n\left(\mathrm{M}_{\hat{x}} - \widehat{\mathrm{M}}^*_{\hat{x}}\right)^2,$$

и

$$-D^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}[(\hat{x}_i - \mathrm{M}_{\hat{x}})^2 - (\widehat{\mathrm{M}}^*_{\hat{x}} - \mathrm{M}_{\hat{x}})^2 = \qquad (5.45)$$

$$= \frac{\delta^2_{\hat{x}}}{n}\left[\sum_{i=1}^{n}\left(\frac{\hat{x}_i - \mathrm{M}_{\hat{x}}}{\delta_{\hat{x}}}\right)^2 - \left(\frac{\widehat{\mathrm{M}}^*_{\hat{x}} - \mathrm{M}_{\hat{x}}}{\delta_{\hat{x}}/\sqrt{n}}\right)^2\right].$$

Как показано в курсах вероятностей, случайная величина

$$\hat{\chi}^2_{(n)} = \sum_{i=1}^{n}\left(\frac{\hat{z}_1 - \mathrm{M}_{\hat{z}}}{\delta_{\hat{z}}}\right)^2, \qquad (5.46)$$

где $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n$ - независимые случайные величины, распределенные нормально и одинаково с $\hat{z}$, распределена по закону χ с n степенями свободы. Таким образом, первое слагаемое в квадратных скобках в выражении (5.45) равно $\hat{\chi}^2_{(n)}$.

Во втором слагаемом положим $\widehat{\mathrm{M}}^*_{\hat{x}} = \hat{z}$. Тогда, учитывая что $\mathrm{M}_{\hat{z}} = \mathrm{M}[\widehat{\mathrm{M}}^*_{\hat{x}}] = \mathrm{M}_{\hat{x}}$, а $\delta_{\hat{z}} = \delta[\widehat{\mathrm{M}}^*_{\hat{x}}] = \delta_{\hat{x}}/\sqrt{n}$, найдем

$$\left(\frac{\widehat{\mathrm{M}}^*_{\hat{x}} - \mathrm{M}_{\hat{x}}}{\delta_{\hat{x}}/\sqrt{n}}\right)^2 = \left(\frac{\hat{z} - \mathrm{M}_{\hat{z}}}{\delta_{\hat{z}}}\right)^2 = \hat{\chi}^2_{(1)}, \qquad (5.47)$$

следовательно,

$$D^*_{\hat{x}} = \frac{\delta^2_{\hat{x}}}{n}(\hat{\chi}^2_{(n)} - \hat{\chi}^2_{(1)}) = \frac{\delta^2_{\hat{x}}}{n}\hat{\chi}^2_{(n)} \qquad (5.48)$$

Но, как известно, $\mathrm{M}[\hat{\chi}^2_{(n-1)}] = n - 1$. Поэтому

$$\mathrm{M}[D^*_{\hat{x}}] = \frac{n-1}{n}D_{\hat{x}}. \qquad (5.49)$$

Таким образом, оценки (5.44) смещены относительно $D_{\hat{x}}$. Несмещенность оценок $\hat{\mathrm{x}}$ обеспечивается использованием статистики

$$D^*_{\hat{x}} = \frac{1}{n-1}\sum_{i=1}^{n}[(\hat{x}_i - \widehat{\mathrm{M}}^*_{\hat{x}})^2 = \frac{\delta^2_{\hat{x}}}{n-1}\hat{\chi}^2_{(n-1)}. \qquad (5.50)$$

Перейдем к анализу состоятельности и эффективности оценок. Поскольку $\mathrm{D}[\hat{\chi}^2_{(n-1)}] = 2(n-1)$, то

$$\mathrm{D}[D^*_{\hat{x}}] = \frac{D^2_{\hat{x}}}{(n-1)^2}2(n-1) = \frac{2D^2_{\hat{x}}}{n-1} \qquad (5.51)$$

стремится к нулю при n → ∞ и оценки, вытекающие из (5.50), асимптотически эффективны и состоятельны. Для оценивания среднего квадратического отклонения $\hat{x}$ (при том же условии нормальности распределения $\hat{x}$) естественно воспользоваться статистикой

$$\hat{\delta}^*_{\hat{x}} = \sqrt{D^*_{\hat{x}}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[(\hat{x}_i - \widehat{M}^*_{\hat{x}})^2}\,\frac{\delta_{\hat{x}}}{n-1}\hat{\chi}_{(n-1)}. \qquad (5.52)$$

Но

$$M[\hat{\chi}_{(n-1)}] = \sqrt{2}\,\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}, \qquad (5.53)$$

где Γ(m) – гамма-функция от m.

Следовательно, $M[\hat{\delta}^*_{\hat{x}}] = \sqrt{\frac{2}{n-1}}\,\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\,\delta_{\hat{x}}$, и оценки $\delta_{\hat{x}}$, определяемые (5.52), смещены.

Несмещенность оценок обеспечивается статистикой

$$\hat{\delta}^*_{\hat{x}} = K_{n-1}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{x}_i - \widehat{M}^*_{\hat{x}})^2} \qquad (5.54)$$

где

$$K_{n-1} = \sqrt{\frac{n-1}{2}}\,\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}. \qquad (5.55)$$

Коэффициент K_{n-1} при не очень малых n весьма близок к единице, при n = 5 $K_{n-1} = 1{,}06$, при n = 10 $K_{n-1} = 1{,}09$, при n = 20 $K_{n-1} = 1{,}01$ и т. д. поэтому при практических расчетах обычно принимают $K_{n-1} \approx 1$ и тогда

$$\hat{\delta}^*_{\hat{x}} = \sqrt{D^*_{\hat{x}}}. \qquad (5.56)$$

Как следует из анализа, выполненного для $D^*_{\hat{x}}$, оценки $\hat{\delta}^*_{\hat{x}}$, вытекающие из (5.54), состоятельны и асимптотически эффективны.

Можно показать, что приведенные оценки $D^*_{\hat{x}}$ и $\hat{\delta}^*_{\hat{x}}$ являются подходящими значениями $D_{\hat{x}}$, $\delta_{\hat{x}}$ и при распределении $\hat{x}$, отличном от нормального. Однако предположение о нормальности распределения $\hat{x}$ значительно упрощает анализ качества оценивания $D_{\hat{x}}$ и $\delta_{\hat{x}}$, к которому мы сейчас перейдем.

Для выборок большого объема при сделанных предположениях распределение $D^*_{\hat{x}}$ близко к нормальному, следовательно,

$$\beta = P(|D^*_{\hat{x}} - D_{\hat{x}}| \le \varepsilon) = 2\varphi_0\left(\frac{\varepsilon}{\delta_{D^*_{\hat{x}}}}\right) = 2\varphi_0\left(\frac{\varepsilon\sqrt{n-1}}{D_{\hat{x}}\sqrt{2}}\right) \approx$$
$$\approx 2\varphi_0\left(\frac{\varepsilon\sqrt{n-1}}{D^*_{\hat{x}}\sqrt{2}}\right) \qquad (5.57)$$

При той же доверительной вероятности β границы доверительного интервала для среднего квадратического отклонения $\delta_{\hat{x}}$ $\delta_{\hat{x}}$ - ε_1 и $\delta_{\hat{x}}$ + ε_2 можно определить через доверительные границы для $D_{\hat{x}}$, найденные из (5.57):

$$\delta^*_{\hat{x}} - \varepsilon_1 = \sqrt{D^*_{\hat{x}} - \varepsilon}, \qquad (5.58)$$

$$\delta^*_{\hat{x}} + \varepsilon_2 = \sqrt{D^*_{\hat{x}} + \varepsilon}. \quad (5.59)$$

Тогда

$$\beta = \mathrm{P}\left(\sqrt{\widehat{D}^*_{\hat{x}} - \varepsilon} \le \delta_{\hat{x}} \le \sqrt{\widehat{D}^*_{\hat{x}} + \varepsilon}\right), \quad (5.60)$$

причем построенный доверительный интервал $\left[\sqrt{D^*_{\hat{x}} - \varepsilon}, \sqrt{D^*_{\hat{x}} + \varepsilon}\right]$ не симметричен относительно $\delta^*_{\hat{x}}$. Соотношение (5.57) и (5.60) позволяют решить все основные задачи анализа качества оценивания $D_{\hat{x}}$ и $\delta_{\hat{x}}$ при большом объеме выборок.

Пример 7. По данным табл. 5.1 (пример 4, подразд. 5.5) требуется оценить дисперсию $D_{\hat{T}}$ и среднее квадратическое отклонение $\delta_{\hat{T}}$ средней температуры воздуха в Санкт-Петербурге в марте $\hat{T}$, считая распределение $\hat{T}$ приближено нормальным. По формуле (5.50) найдем оценку дисперсии $\hat{T}$: $D^*_{\hat{T}} = \frac{1}{54}\sum_{i=1}^{55}\left(T_i - \frac{1}{54}\sum_{i=1}^{55} T_i\right)^2 = 7{,}24$ (℃), и по формуле оценку среднеквадратического отклонения $\hat{T}$: $\delta^*_{\hat{T}} = \sqrt{D^*_{\hat{T}}} = 2{,}69$ ℃. Для анализа качества оценивания дисперсии воспользуемся формулой (5.57):

$$\beta \approx 2\varphi_0\left(\frac{\varepsilon\sqrt{54}}{7{,}24\sqrt{2}}\right).$$

Найдем, например, границы доверительного интервала при $\beta = 0{,}95$. По таблице 1 Приложения для $\varphi_0(x) = 0{,}475$ x = 1,96 и следовательно $\left(\frac{\varepsilon\sqrt{27}}{7{,}24}\right) = 1{,}96$, откуда $\varepsilon = 2{,}73$(℃)². При той же доверительной вероятности доверительные границы для $\delta_{\hat{T}}$ равны: $\sqrt{4{,}51} = 2{,}12$ и $\sqrt{9{,}97} = 3{,}16$℃. При малом объеме выборок распределение $\widehat{D}*$ существенно отличается от нормального , и оценивание дисперсии на основании (5.57) может привести к большим погрешностям. В этих ситуациях руководствуются следующими соображениями. Вернемся к исходной формуле (5.16), которая применительно к оцениванию дисперсии приобретает вид

$$\beta = \mathrm{P}(\widehat{D}_1 \le D_{\hat{x}} \le \widehat{D}_2). \quad (5.61)$$

Пусть, для получения точечных оценок D используется статистика $\widehat{D}*$. Тогда под β можно понимать вероятность попадания $\widehat{D}*$ в интервал $[D_1;\ D_2]$, то есть

$$\beta = \mathrm{P}(D_1 \le \widehat{D}^*_{\hat{x}} \le D_2). \quad (5.62)$$

Согласно (5.50)

$$\widehat{D}^*_{\hat{x}} = \frac{D_{\hat{x}}}{n-1}\hat{\chi}^2_{(n-1)} \quad (5.63)$$

следовательно, доверительные границы для $D_{\hat{x}}$ можно выразить через доверительные границы для $\chi^*_{(n-1)}$. Выберем последние таким образом, чтобы вероятности выхода $\chi^*_{(n-1)}$ за пределы доверительного интервала вправо и влево были одинаковыми (рис. 5.1).

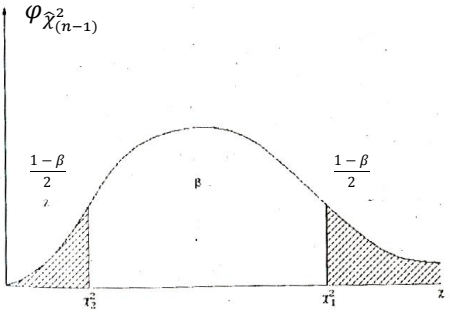


Рис. 5.1

При таком выборе

$$F_{\hat{\chi}^2_{(n-1)}}(\chi^2_2) = \frac{1-\beta}{2}, \tag{5.64}$$

$$F_{\hat{\chi}^2_{(n-1)}}(\chi^2_1) = 1 - \frac{1-\beta}{2} = \frac{1+\beta}{2}, \tag{5.65}$$

и доверительные границы для $\chi^2_{(n-1)}$ равны:

$$\chi^2_2 = F^{-1}_{\hat{\chi}^2_{(n-1)}}\left(\frac{1-\beta}{2}\right), \tag{5.66}$$

$$\chi^2_1 = F^{-1}_{\hat{\chi}^2_{(n-1)}}\left(\frac{1+\beta}{2}\right). \tag{5.67}$$

Но в силу (5.63), при $\hat{\chi}^2_{(n-1)} \geq \chi^2_2$ $\hat{D}^*_{\hat{x}} \geq \frac{D_{\hat{x}}}{n-1}\chi^2_2$, а при $\hat{\chi}^2_{(n-1)} \leq \chi^2_1$ $\hat{D}^*_{\hat{x}} \leq \frac{D_{\hat{x}}}{n-1}\chi^2_1$. Таким образом, для $D_{\hat{x}}$ доверительными границами при заданном β должны считаться:

$$\begin{aligned} D_1 &= \frac{D^*_{\hat{x}}}{\chi^2_1}(n-1) \text{ – «правая» граница;} \\ D_2 &= \frac{D^*_{\hat{x}}}{\chi^2_2}(n-1) \text{ – «левая» граница.} \end{aligned} \tag{5.68}$$

Доверительные границы для среднего квадратического отклонения находятся из равенства $\delta_{\hat{x}} = \sqrt{D_{\hat{x}}}$:

$$\delta_1 = \sqrt{\frac{D^*_{\hat{x}}}{\chi^2_1}(n-1)}\,;\ \delta_2 = \sqrt{\frac{D^*_{\hat{x}}}{\chi^2_2}(n-1)}\,; \tag{5.69}$$

Пример 8. По данным табл. 5.2 (пример 6, подразд. 5.6) построить доверительные интервалы для $D_{\Delta\hat{p}}$ и $\delta_{\Delta\hat{p}}$ при $\beta = 0{,}90$.

В примере 6 (подразд 5.6) получено:

$\delta_{\Delta\hat{p}}$ = 5,25 мбар, $D_{\Delta\hat{p}}$ = 27,56 мбар.

По условиям задачи $\frac{1+\beta}{2}$ = 0,95, $\frac{1-\beta}{2}$ = 0,05, n – 1 9. По табл. 3 Приложения, приняв θ_1 = 1 – 0,95 = 0,05, θ_2 = 1 – 0,05 = 0,95, найдем χ_1^2 = 16,9; χ_2^2 = 3,32 согласно (5.68): $D_1 = \frac{27,56}{16,9}9$ = 14,68, $D_2 = \frac{27,56}{3,32}9$ = 74,71 мбар². Доверительные границы $\delta_{\Delta\hat{p}}$ соответственно равны: $\delta_1 = \sqrt{D_1}$ = 3,83, $\delta_2 = \sqrt{D_2}$ = 8,64 мбар.

Обращает на себя внимание несимметричность интервалов относительно D* и δ*. Так, при $D^*_{\Delta\hat{p}}$ = 27,56 мбар² середине доверительного интервала соответствует D = 44,70 мбар². Интересно сопоставить полученные значения доверительных границ с результатами оценивания при аппроксимации распределения $D^*_{\Delta\hat{p}}$ нормальным законом.

По формуле (5.57) 0,90 = $2\varphi_0\left(\frac{\varepsilon\sqrt{9}}{27,56\sqrt{2}}\right)$. По табл. 1 Приложения находим $\frac{\varepsilon 3}{27,56\sqrt{2}} \approx$ 1,65. Следовательно $\varepsilon = \frac{1,65}{3}27,56\sqrt{2}$ = 21,44 мбар² и D_1 = 6,12, D_2 = 49,00 мбар² (а не 14,68 и 74,71 мбар², как получено выше).

5.8 Оценивание коэффициента корреляции между случайными величинами

Ранее уже рассматривались задачи оценивания числовых характеристик одномерных распределений. При статистическом анализе многомерных распределений – распределений случайных векторов – возникает необходимость в оценивании дополнительных характеристик, наиболее важной из которых является коэффициент корреляции. Как известно из теории вероятностей, коэффициент корреляции $r_{\hat{x},\hat{y}}$ между случайными величинами $\hat{x}$ и $\hat{y}$ связан с корреляционным моментом $к_{\hat{x},\hat{y}}$ и средними квадратическими отклонениями $\delta_{\hat{x}}$ и $\delta_{\hat{y}}$ выражением:

$$r_{\hat{x},\hat{y}} = \frac{к_{\hat{x},\hat{y}}}{\delta_{\hat{x}}\delta_{\hat{y}}}, \qquad (5.70)$$

где $к_{\hat{x},\hat{y}}$ = M[($\hat{x}$ − $M_{\hat{x}}$)($\hat{y}$ - $M_{\hat{y}}$)].

Оценки введенных характеристик должны быть получены на атериале независимых наблюдений, результаты которых представлены в виде последовательности пар чисел x_1, y_1 (табл. 5.3).

Таблица 5.3

i	1	2	…	i	…	N
x_i	x_1	x_2	…	x_i	…	x_n
y_i	y_1	y_2	…	y_i	…	y_n

Вопросы оценивания математических ожиданий, дисперсий и средних квадратических отклонений случайных величин $\hat{x}$ и $\hat{y}$ рассмотрены ранее. Поэтому ограничимся задачами оценивания корреляционного момента $К_{\hat{x},\hat{y}}$ и коэффициента корреляции $r_{\hat{x},\hat{y}}$. Нетрудно видеть, что РМ – оценка $K_{\hat{x},\hat{y}}$ находится по формуле:

$$K^*_{\hat{x},\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{x}_i - \mathrm{M}^*_{\hat{x}})(y_i - \mathrm{M}^*_{\hat{y}}) \qquad (5.71)$$

Такие оценки состоятельны, асимптотически эффективны, но смещены относительно $K_{\hat{x},\hat{y}}$. Можно показать, аналогично оцениванию дисперсии случайных величин в п. 5.7, что оценки

$$K^*_{\hat{x},\hat{y}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{x}_i - \mathrm{M}^*_{\hat{x}})(y_i - \mathrm{M}^*_{\hat{y}}) \qquad (5.72)$$

не смещены и являются подходящими значениями $к_{\hat{x},\hat{y}}$.

РМ – оценки коэффициента корреляции $r_{\hat{x},\hat{y}}$ даются формулой

$$r_{\hat{x},\hat{y}} = \frac{K^*_{\hat{x},\hat{y}}}{\delta^*_{\hat{x}}\delta^*_{\hat{y}}}. \qquad (5.73)$$

По аналогии с ранее рассмотренными характеристиками можно ожидать, что распределение оценок коэффициента корреляции с увеличением n сходится к нормальному. Оказывается, что предположение о близости распределения $\hat{r}^*$ нормальному практически с удовлетворительной точностью выполняется при n > 30, причем параметры аппроксимирующего нормального закона равны $\mathrm{M}_{\hat{r}^*}$ = r и $\delta_{\hat{r}^*} = \frac{\sqrt{1-r^2}}{n}$. Поэтому для больших n выполняется приближенное равенство (5.22), то есть:

$$\beta \approx 2\varphi_0\left(\frac{\varepsilon}{\delta_{\hat{r}^*}}\right) = 2\varphi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{1-r^2}}\right) \approx 2\varphi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{1-r^2}}\right) \qquad (5.74)$$

При малых n приближение (5.74) приводит к неудовлетворительным результатам, и целесообразно аппроксимировать нормальным законам не распределение $\hat{r}^*$, а распределение случайной величины $\hat{z}^*$, связанной с $\hat{r}^*$ формулой

$$\hat{z}^* = \frac{1}{2}\ln\frac{1+\hat{r}^*}{1-\hat{r}^*}. \qquad (5.75)$$

Как показано Фишером,

$$\mathrm{M}_{\hat{z}^*} \approx \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{r}{2(n-1)}, \qquad (5.76)$$

$$\delta_{\hat{z}^*} \approx \frac{1}{\sqrt{n-3}}, \qquad (5.77)$$

причем распределение $\hat{z}^*$ сходится к нормальному значительно быстрее, чем распределение $\hat{r}^*$. Поскольку при не очень малых n, как следует из (5.76) $\mathrm{M}_{\hat{z}^*}$

$\approx \frac{1}{2} \ln \frac{1+r}{1-r}$ = z, то есть оценки z* практически не смещены, справедливо приближенное равенство (5.78):

$$\beta = \mathrm{P}(|\mathrm{z} * - z| \leq \varepsilon) = 2\varphi_0 \left(\frac{\varepsilon}{\hat{\delta}^*}\right) = 2\varphi_0 \left(\varepsilon\sqrt{n-3}\right). \qquad (5.78)$$

На основании (5.78) при заданной доверительной вероятности β по таблице интеграла вероятностей может быть найдено значение $\varepsilon\sqrt{n-3}$, соответствующее $\varphi_0 = \frac{\beta}{2}$, и определены доверительные границы для z: z_1 = z – ε = $\mathrm{M}_{\hat{z}^*}$ - ε и z_2 = z + ε = $\mathrm{M}_{\hat{z}^*}$ + ε. Так как из (5.75)

$$\mathrm{r} = \frac{e^{2z}-1}{e^{2z}+1}, \qquad (5.79)$$

Доверительные границы для r определяются подстановкой в (5.79) z z_1 и z z_2.

Пример 9. По данным табл. 5.2 (обе выборки) оценить коэффициент корреляции между температурой воздуха у поверхности земли ($\hat{\mathrm{x}}_1$) и на изобарической поверхности 850 гПа ($\hat{\mathrm{x}}_2$) в пункте зондирования.

По формуле (5.31), (5.53) найдем: $\mathrm{M}^*_{\hat{\mathrm{x}}_1}$ = 9,883 ℃, $\mathrm{M}^*_{\hat{\mathrm{x}}_2}$ = 8,267 ℃, $\delta^*_{\hat{\mathrm{x}}_1}$ = 5,459 ℃, $\delta^*_{\hat{\mathrm{x}}_2}$ = 5,217 ℃. Расчеты по формулам (5.72) и (5.73) дают следующую оценку коэффициента корреляции между $\hat{\mathrm{x}}_1$ и $\hat{\mathrm{x}}_2$: $r^*_{\hat{\mathrm{x}}_1,\hat{\mathrm{x}}_2}$ = 0,849.

Для построения доверительного интервала воспользуемся преобразованием Фишера. Сделаем это, например, для β = 0,85. По таблице интеграла вероятностей в этом случае найдем, что аргумент $\varepsilon\sqrt{n-3}$ в (5.78) равен 1,44 следовательно, (при n = 60) $\varepsilon = \frac{1{,}44}{\sqrt{57}}$ = 0,191.

Но по (5.75) z* = 1,253.

Таким образом, доверительные границы для z:

z_1 = 1,253 – 0,191 = 1,062 и z_2 = 1,253 + 0,194 = 1,444.

Подставим полученные значения в формулу (4.79), найдем доверительные границы для $r_{\hat{\mathrm{x}}_1,\hat{\mathrm{x}}_2}$: r_1 = 0,786 и r_2 = 0,895. Если воспользоваться менее строгой формулой (5.74), то получим ε = 1,44$\sqrt{\frac{1-0{,}849^2}{60}}$ = 0,098 следовательно, r_1 = 0,751 и r_2 = 0,947.

6 СТАТИСТИЧЕСКИЙ АНАЛИЗ СВЯЗЕЙ

6.1 Предварительные сведения

В метеорологической практике широко используются стохастические модели связей между различными характеристиками состояния атмосферы. Методам построения стохастических моделей связей между характеристиками изучаемых объектов на основе статистической обработки результатов наблюдений посвящен один из основных разделов прикладной статистики – статистический анализ связей.

В данном случае для указанных характеристик будем пользоваться общим названием – переменные, и обозначениями: $\hat{x}_0$, $\hat{x}_1$, …, $\hat{x}_n$. Предполагается, что в распоряжении статистики имеется случайная выборка из генеральной совокупности, содержащая m реализаций случайного вектора $\hat{\chi}_{<n+1>} = \langle \hat{x}_0, \hat{x}_1, \ldots, \hat{x}_n \rangle$.

Принято выделять два основных типа связей между рассматриваемыми переменными: взаимные статистические связи между компонентами вектора $\hat{\chi}_{<n+1>}$ и статистическую зависимость какой-либо одной или нескольких компонент вектора $\hat{\chi}_{<n+1>}$ от остальных переменных.

Модели связей первого типа находят применение при решении задач классификации объектов и преобразования системы переменных. Задачи классификации в оперативной метеорологической практике встречаются редко и в учебнике не рассматриваются.

Задачи преобразования системы переменных возникают, как правило, при статистическом анализе связей между большим количеством переменных в силу ограниченности объемов метеорологических архивов. В этих ситуациях целесообразен переход к новой системе компонент вектора $\hat{\chi}_{<n+1>}$, построенной таким образом, чтобы основная информация о реализации этого вектора была бы сосредоточена всего в нескольких переменных. Особенно полезным указанное преобразование бывает при разработке методов метеорологических прогнозов.

Вопросам «сжатия» информации путем преобразования системы переменных посвящены подразд. 6.4 компонентный n-факторный анализ. Пример использования компонентного анализа в прогностических целях приводится в подразд. 6.2.9.

Модели связей второго типа предназначена для составления прогнозов реализации переменной $\hat{x}_0$ по результатам наблюдений за остальными компонентами вектора $\hat{\chi}_{<n+1>}$, контроля этих результатов и их пространственной интерполяции. Теория и прикладные вопросы исследования таких связей в ситуациях, когда все компоненты вектора $\hat{\chi}_{<n+1>}$ - количественные характеристики, составляют содержание регрессионного анализа. Методами регрессионного анализа

могут решаться, например, задачи интерполяции результатов наблюдений в узлы регулярной сетки точек «восстановления» метеорологических полей за границами территории, освященной наблюдениями, прогноза ночного понижения температуры воздуха по наблюдениям в вечерние часы за температурой и влажностью воздуха, скоростью ветра и количеством облачности различных ярусов. Вопросами регрессионного анализа посвящен подразд. 6.2.

Ситуации, когда x_0 - количественная характеристика, а все остальные переменные – качественные, исследуются методами дисперсионного анализа. Такая ситуация в метеорологической практике возникает, например, при анализе зависимости количества осадков от их вида, типа измерительного прибора и расположения метеостанции. Методы дисперсионного анализа рассматриваются в подразд. 6.3.

В ситуациях, когда x_0 по-прежнему количественная характеристика среди остальных переменных есть и количественные и качественные характеристики, основным методом исследования связей является ковариационный анализ, которому посвящен подразд. 6.4. методы ковариационного анализа могут быть использованы, например, при решении упоминавшейся задачи прогноза ночного понижения температуры воздуха, если в число наблюдаемых характеристик дополнительно включены форма облаков, тип синоптической обстановки и т. п.

Для изучения зависимостей качественных характеристик от остальных компонент вектора $\hat{\chi}_{<n+1>}$ используются методы дискриминантного анализа, рассматриваемые в подразд. 6.4. Методы дискриминантного анализа широко используются в метеорологической практике, например, при составлении альтернативных прогнозов погодных явлений (гроз, туманов и др.).

Построение стохастических моделей, как правило, начинается с оценивания тесноты рассматриваемых связей, выполняемого методами корреляционного анализа. Некоторые прикладные аспекты корреляционного анализа обсуждаются в подразд. 6.2.

Более полное изложение задач и методов корреляционного анализа читатель может найти в рекомендованной литературе.

В процессе статистического анализа связей необходимо последовательное решение следующих задач.

Основываясь на целях исследования, следует определить перечень рассматриваемых переменных и критерии качества разрабатываемых моделей.

Установить наличие и тесноту связей между назначенными переменными.

Выбрать тип математической модели связей.

Оценить значения параметров выбранной модели.

Проанализировав качество модели и, если оно окажется неудовлетворительным, внести необходимые изменения в перечень переменных и структуры модели.

6.2 Регрессионный анализ

6.2.1 Постановка задачи регрессионного анализа

При разработке вероятностных моделей атмосферных процессов часто возникает необходимость в анализе зависимости некоторой случайной величины $\hat{x}_0$ от случайных величин $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ (случайного вектора $\hat{\chi}_{<n+1>}$).

Случайные величины $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ в метеорологической литературе принято называть предикторами (иногда – регрессорами или независимыми переменными), а случайную величину $\hat{x}_0$ - предиктантом (иногда – зависимой переменной).

При регрессионном анализе случайная величина $\hat{x}_0$ представляется в виде суммы функции от случайного вектора $\hat{\chi}_{<n+1>}$, называемой регрессией $\hat{x}_0$ на $\hat{\chi}_{<n+1>}$, и случайной величины $\hat{\varepsilon}$ - ошибки (невязки) регрессии:

$$\chi_{<n>} = f(\chi_{<n>}) + \hat{\varepsilon} \qquad (6.1)$$

В зависимости от числа предикторов n принято различать однокомпонентную (n = 1), двухкомпонентную (n = 2) и многокомпонентную или множественную (n > 2) регрессии. Если функция f линейна относительно предикторов, регрессию принято называть линейной. Иногда при регрессионном анализе в качестве предикторов $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ используются не непосредственно результаты наблюдений $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_k$, а некоторые функции от этих результатов: $\hat{x}_1 = \hat{x}_1 (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_k)$ и т. п. если при этом $f(\chi_{<n>})$ так же линейно относительно $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$, говорят, что регрессия строится в классе обобщенных линейных функций.

Основной задачей регрессионного анализа является нахождение такой функции $f(\chi_{<n>})$, которая обеспечивала бы понимаемую в вероятностном смысле максимальную близость ее значений к значениям предиктанта $\hat{x}_0$. Регрессия

$$\hat{\tilde{x}}_0 = f(\chi_{<n>}), \qquad (6.2)$$

обеспечивающая минимум (на множестве функций f) математического ожидания квадрата ошибки регрессии

$$M\left[\left(\hat{\tilde{x}}_0 - \hat{x}_0\right)^2\right] = M[\hat{\varepsilon}^2] = \min_{(f)}, \qquad (6.3)$$

называется среднеквадратической регрессией $\hat{x}_0$ на $\chi_{<n>}$.

В данном параграфе основное внимание будет уделено линейной среднеквадратической регрессии, наиболее часто используемой в метеорологической практике, то есть регрессии вида

$$\hat{\tilde{x}}_0 = c_0 + c_1 \hat{x}_1 + c_2 \hat{x}_2 + \ldots + c_n \hat{x}_n, \qquad (6.4)$$

Где коэффициенты c_0, c_1, …, c_n (коэффициенты регрессии) определены в соответствии с формулой (6.3). В подразд. 6.2.2 – 6.2.4 задачи регрессионного анализа рассматриваются в предположении, что apriori известно совместное распределение предиктанта $\hat{x}_0$ и вектора – предиктора $\chi_{<n>}$, подразд. 6.2.5 посвящен выборочному регрессионному анализу (т. е., по сути говоря, вопросам статистического оценивания регрессии). Проблема формирования оптимального перечня предикторов рассматривается в гл. 8.

6.2.2 Однокомпонентный анализ

Ознакомление с целями и методами регрессионного анализа начнем с рассмотрения следующих задач. Основываясь только на информации о распределений «новогодних» значений температуры воздуха в Санкт-Петербурге $\hat{T}$, требуется: 1). Составить прогноз этой температуры в ночь на 1 января 2000-го года (T_{2000}) и 2). Определить температуру в Санкт-Петербурге в ночь на 1 января 1500 года (T_{1600}).

Что характерно для подобного рода задач? Требуется по известному закону распределения случайной величины $\hat{x}_0$ указать значение, которое должно принять $\hat{x}_0$ в конкретном испытании («составить прогноз») этого значения). Каким правилом следует руководствоваться при решении таких задач?

Прежде всего отметим, что, поскольку каждый раз при прогнозе значений $\hat{x}_0$ используется одно и то же безусловное распределение $\hat{x}_0$, во всех прогнозах естественно указывать одно и то же безусловное распределение $\hat{x}_0$, во всех прогнозах естественно указывать одно и то же «наилучшее» приближенное значение $\hat{x}_0$. Обозначим это значение через $\tilde{x}_0$. Выбор $\tilde{x}_0$, вообще говоря, определяется требованиями к распределению ошибок прогноза $\hat{\varepsilon} = \tilde{x}_0 - \hat{x}_0$ со стороны потребителя метеорологической информации.

Общие методы такого выбора были рассмотрены в гл. 2. В данном подразделе ограничимся анализом типичных для метеорологической практики ситуаций, когда лучшим признаются прогнозы, обеспечивающие минимум математического ожидания квадрата ошибки прогноза $\hat{\varepsilon}$, то есть удовлетворяющие условию (6.3).

Нетрудно показать, что это условие выполняется, если принять $M_{\hat{x}_0}$. Действительно, для определения минимума $M[\hat{\varepsilon}^2]$ на множестве значений $\tilde{x}_0$ следует положить:

$$\frac{\partial}{\partial \tilde{x}_0} M[\hat{\varepsilon}^2] = \frac{\partial}{\partial \tilde{x}_0} M[(\tilde{x}_0 - \hat{x}_0)^2] = 2\,(\tilde{x}_0 - M_{\hat{x}_0}) = 0,$$

и, следовательно,

$$\tilde{x}_0 = M_{\hat{x}_0}. \quad (6.5)$$

Поскольку $\frac{\partial^2}{\partial \tilde{x}_0^2} M[\varepsilon^2] = 2 > 0$, условие (6.5) соответствует минимуму $M[\hat{\varepsilon}_1^2]$.

Таким образом, регрессия (6.4) в рассмотренной ситуации принимает вид (6.5), причем качество прогнозов характеризуется величиной дисперсии $\hat{x}_0$

$$M[\hat{\varepsilon}_0^2] = M[\varepsilon^2] = M[(\tilde{x}_0 - \hat{x}_0)^2] = M\left[\left(M_{\hat{x}_0} - \hat{x}_0\right)^2\right] = D_{\hat{x}_0}. \quad (6.6)$$

Перейдем к рассмотрению более сложной задачи. Пусть известно совместное распределение температуры воздуха в моменты наибольшего перегрева $\hat{T}_{max}$ и восхода солнца $\hat{T}_g$ для некоторого географического пункта в определенной календарной даты (например, 1 июля), и пусть осуществилось значение $\hat{T}_g$, равное T_g. Какое значение следует указать в прогнозе $\hat{T}_{max}$ на текущие сутки, основываясь на указанной информации?

В общем виде задачи такого типа формулируются следующим образом. Задан закон совместного распределения случайных величин $\hat{x}_0$ и $\hat{x}_1$, и известна реализация x_1 случайной величины $\hat{x}_1$. Какое значение $\hat{x}_0$, соответствующее этой реализации, следует указать в прогнозе величины $\hat{x}_0$?

В отличие от предыдущих примеров значения $\tilde{x}_0$, указываемые в прогнозах $\hat{x}_0$, вообще говоря, должны быть различными при различных реализациях $\hat{x}_1$, причем для выполнения условия (6.3) очевидно необходимо при реализации $\hat{x}_1 = x_1$ прогнозировать осуществление значения $\hat{x}_0$, равного условному математическому ожиданию $\hat{x}_0$ при $\hat{x}_1 = x_1$, т. е. принять

$$\tilde{x}_0(x_1) = M[\hat{x}_0/\hat{x}_1 = x_1]. \quad (6.7)$$

Качество таких прогнозов характеризуется величиной условной дисперсии $\hat{x}_0$ при $\hat{x}_1 = x_1$:

$M[\hat{\varepsilon}^2(x_1)] = D[\hat{x}_0/\hat{x}_1 = x_1]$,

а среднее по всем реализациям $\hat{x}_1$ качество прогнозов – величиной

$M[\hat{\varepsilon}_1^2] = M[\hat{\varepsilon}^2] = M_{\hat{x}_1}[D_{\hat{x}_0/\hat{x}_1}]$.

Разность $M[\hat{\varepsilon}_0^2] - M[\hat{\varepsilon}_1^2]$ характеризует повышение качества рассмотренных прогнозов за счет учета связи $\hat{x}_0$ с $\hat{x}_1$ и в этом смысле может считаться мерой тесноты статистической связи между случайными величинами $\hat{x}_0$ и $\hat{x}_1$. Обычно эта разность «обезразмеривается» делением ее на $M[\hat{\varepsilon}_0^2]$, и рассматривается отношение

$$\frac{M[\hat{\varepsilon}_0^2] - M[\hat{\varepsilon}_1^2]}{M[\hat{\varepsilon}_0^2]} = \frac{D_{\hat{x}_0} - M[\hat{D}_{\hat{x}_0/\hat{x}_1}]}{D_{\hat{x}_0}} = \eta_{0,1}^2. \quad (6.8)$$

В формуле (6.8)

$D_{\hat{x}_0}$ - полная дисперсия случайной величины $\hat{x}_0$;

$M[\hat{D}_{\hat{x}_0/\hat{x}_1}]$ - остаточная дисперсия $\hat{x}_0$;

$D_{\hat{x}_0} - \mathrm{M}\,[\widehat{D}_{\hat{x}_0/\hat{x}_1}]$ - снятая дисперсия $\hat{x}_0$;

$\eta_{0,1}$ - индекс корреляции, характеризующий тесноту статистической связи между $\hat{x}_0$ и $\hat{x}_1$ и меняющийся от 0 при статистической независимости $\hat{x}_0$ и $\hat{x}_1$ до 1 при функциональной связи между ними.

Для регулярного составления прогнозов $\hat{x}_0$ по правилу (6.7) необходимо определить зависимость $\mathrm{M}\,[\hat{x}_0/\hat{x}_1 = x_1]$ от $\hat{x}_1$, то есть построить среднеквадратическую регрессию $\hat{x}_0$ на $\hat{x}_1$:

$$\hat{\hat{x}}_0 = \mathrm{M}\,[\hat{x}_0/\hat{x}_1] = f(\hat{x}_1). \tag{6.9}$$

Как отмечалось в подразд. 6. 2. 1, функция $f(\hat{x}_1)$ обычно ищется в классе линейных (или обобщенных линейных) функций. Покажем, что при нормальном совместном распределении $\hat{x}_0$ и $\hat{x}_1$ «лучшая» $f(\hat{x}_1)$ строго линейна.

Плотность совместного нормального распределения случайных величин $\hat{x}_0$ и $\hat{x}_1$ выражается известной формулой:

$$\varphi_{\hat{x}_0,\hat{x}_1}(x_0, x_1) = \frac{1}{2\pi\delta_{\hat{x}_0}\delta_{\hat{x}_1}\sqrt{1-r^2_{\hat{x}_0,\hat{x}_1}}} \times \\ \times e^{\frac{1}{2\left(1-r^2_{\hat{x}_0,\hat{x}_1}\right)}\left[\frac{\left(x_0-\mathrm{M}_{\hat{x}_0}\right)^2}{\delta^2_{\hat{x}_0}} \frac{2r_{\hat{x}_0,\hat{x}_1}\left(x_0-\mathrm{M}_{\hat{x}_0}\right)\left(x_1-\mathrm{M}_{\hat{x}_1}\right)}{\delta_{\hat{x}_0}\delta_{\hat{x}_1}} \frac{\left(x_1-\mathrm{M}_{\hat{x}_1}\right)^2}{\delta^2_{\hat{x}_1}}\right]}, \tag{6.10}$$

где $r_{\hat{x}_0,\hat{x}_1}$ - коэффициент корреляции между x_0 и x_1, а плотность безусловного распределения x_1 выражается формулой

$$\varphi_{\hat{x}_1}(x_1) = \frac{1}{\sqrt{2\pi\delta_{\hat{x}_1}}} e^{\frac{\left(\hat{x}_1-\mathrm{M}_{\hat{x}_1}\right)^2}{2\delta^2_{\hat{x}_1}}}. \tag{6.11}$$

Но

$$\varphi_{\hat{x}_0/\hat{x}_1} = \frac{\varphi_{\hat{x}_0,\hat{x}_1}}{\varphi_{\hat{x}_1}},$$

поэтому с учетом формулы (6.10) и (6.11) получаем

$$\varphi[\hat{x}_0/\hat{x}_1 = x_1] = \frac{1}{\delta_{\hat{x}_0}\sqrt{2\pi}\sqrt{1-r^2_{\hat{x}_0,\hat{x}_1}}} \times \\ \times e^{2\delta^2_{\hat{x}_1}\left(\frac{1}{1-r^2_{\hat{x}_0,\hat{x}_1}}\right)\left[x_0-\mathrm{M}_{\hat{x}_0}-r_{\hat{x}_0,\hat{x}_1}\frac{\delta_{\hat{x}_0}}{\delta_{\hat{x}_1}}\left(x-\mathrm{M}_{\hat{x}_1}\right)\right]^2},$$

Следовательно, условное распределение $\hat{x}_0$ подчинено нормальному закону математическим ожиданием

$$\mathrm{M}\,[\hat{x}_0/\hat{x}_1 = x_1] = \mathrm{M}_{\hat{x}_0} + r_{\hat{x}_0,\hat{x}_1}\frac{\delta_{\hat{x}_0}}{\delta_{\hat{x}_1}}\left(x_1 - \mathrm{M}_{\hat{x}_1}\right).$$

Таким образом, при совместном нормальном распределении $\hat{x}_0$ и $\hat{x}_1$ регрессия (6.9) действительно линейна:

$$\hat{\tilde{x}}_0 = M[\hat{x}_0/\hat{x}_1] = c_0 + c_1 \hat{x}_1, \tag{6.12}$$

где $c_1 = r_{\hat{x}_0,\hat{x}_1} \frac{\delta_{\hat{x}_0}}{\delta_{\hat{x}_1}}$, $c_0 = M_{\hat{x}_0} - c_1 M_{\hat{x}_1}$.

Существуют и некоторые другие совместные распределения $\hat{x}_0$ и $\hat{x}_1$ при которых среднеквадратическая регрессия линейна. Однако для большинства метеорологических величин зависимость $M[\hat{x}/\hat{x}_1 = x_1]$ от $\hat{x}_1$ существенно нелинейна. И тем не менее эта зависимость, как правило, аппроксимируется линейной функцией

$$\hat{\tilde{x}}_0 = c_0 + c_1 \hat{x}_1, \tag{6.13}$$

где, вообще говоря, $\tilde{x} \neq M[\hat{x}_0/\hat{x}_1 = x_1]$.

Целесообразность такой аппроксимации определяется несколькими причинами. Во-первых, как отмечалось в подразд. 6.2.1, регрессия может строиться в классе обобщенных линейных функций. Во-вторых, методы линейного регрессионного анализа сравнительно просты и хорошо разработаны. В-третьих, при выборочном анализе качество оценивания регрессии по выборке фиксированного объема быстро убывает при увеличении числа оцениваемых параметров (см. гл. 8), и, с этой точки зрения использование сравнительно малопараметрических линейных функций более предпочтительно.

Коэффициенты уравнения (6.13), удовлетворяющие требованию (6.3), являются решениями системы уравнений

$$\frac{\partial}{\partial c_0} M\left[\left(\hat{\tilde{x}}_0 - \hat{x}_0\right)^2\right] = 2M[c_0 + c_1 \hat{x}_1 - \hat{x}_0] = 0, \tag{6.14}$$

$$\frac{\partial}{\partial c_1} M\left[\left(\hat{\tilde{x}}_0 - \hat{x}_0\right)^2\right] = 2M[\hat{x}_1(c_0 + c_1 \hat{x}_1 - \hat{x}_0)] = 0, \tag{6.15}$$

Из уравнения (6.15) следует:

$$c_0 = M_{\hat{x}} - c_1 M_{\hat{x}_1}. \tag{6.16}$$

Подставив полученное выражение в уравнение (6.15), получим:

$M[\hat{x}_1(c_1 \hat{x}_1' - \hat{x}_0')] = 0,$

где $\hat{x}_0' = \hat{x}_0 - M_{\hat{x}_0}$, $\hat{x}_1' = \hat{x}_1 - M_{\hat{x}_1}$,

или

$M\left[(\hat{x}_1' + M_{\hat{x}_1})(c_1 \hat{x}_1' - \hat{x}_0')\right] = M[\hat{x}_1'(c_1 \hat{x}_1' - \hat{x}_0')] = 0,$

откуда

$$c_1 = \frac{M[\hat{x}_0', \hat{x}_1']}{M[\hat{x}_1'^2]} \tag{6.17}$$

Введем в рассмотрение дисперсии в средние квадратические отклонения

$D_{\hat{x}} = \delta_{\hat{x}}^2 = M\left[\hat{x}_{0_2}'^2\right],$

$D_{\hat{x}} = \delta_{\hat{x}}^2 = M\left[\hat{x}_{0_1}'\right],$

корреляционный момент и коэффициент корреляции

$K_{\hat{x}_0,\hat{x}_1} = M[\hat{x}_0', \hat{x}_1'],$

$r_{\hat{x}_0,\hat{x}_1} = \frac{K_{\hat{x}_0,\hat{x}_1}}{\delta_{\hat{x}_0}\delta_{\hat{x}_1}}$.

С использованием введенных обозначений формула (6.17) приводится к виду

$$c_1 = r_{\hat{x}_0,\hat{x}_1}\frac{\delta_{\hat{x}_0}}{\delta_{\hat{x}_1}}. \quad (6.18)$$

Сравнение формул (6.16) и (6.18) с формулой (6.12) показывает, что коэффициенты «точной» (при нормальной совместном распределении $\hat{x}_0$ и $\hat{x}_1$) и «приближенной» среднеквадратической регрессии находятся по одним и тем же формулам.

Качество регрессии (6.13) будем, как и в предыдущих примерах, характеризовать величиной

$M[\hat{\varepsilon}_{01}^2] = M[\hat{\varepsilon}^2] = M\left[\left(\hat{\hat{x}}_0 - \hat{x}_0\right)^2\right] = M[(c_0 + c_1\hat{x}_1 - \hat{x}_0)^2]$.

с учетом (6.15)

$M[\hat{\varepsilon}_{0-1}^2] = M[(c_1\hat{x}_1' - \hat{x}_0')^2] = c_1^2 D_{\hat{x}_0} + D_{\hat{x}_0} - 2c_1 K_{\hat{x}_0,\hat{x}_1}$,

откуда после подстановки вместо c_1 выражения (6.18) получаем:

$$M[\hat{\varepsilon}_{01}^2] = D_{\hat{x}_0}(1 - r_{\hat{x}_0,\hat{x}_1}^2) \quad (6.19)$$

или, вспомнив (6.7)

$$M[\hat{\varepsilon}_{0-1}^2] = M[\hat{\varepsilon}_0^2](1 - r_{\hat{x}_0,\hat{x}_1}^2). \quad (6.20)$$

Из последней формулы следует

$$r_{\hat{x}_0,\hat{x}_1}^2 = \frac{M[\hat{\varepsilon}_0^2] - M[\hat{\varepsilon}_{01}^2]}{M[\hat{\varepsilon}_0^2]}, \quad (6.21)$$

и, так как при нелинейной зависимости $M[\hat{x}_0/\hat{x}_1 = x_1]$ от x_1 величины $M[\hat{\varepsilon}_{01}^2]$, соответствующая «приближенной» регрессии (6.13), чем для «точной» регрессии (6.12).

$r_{\hat{x},\hat{x}}^2 \le \eta_{0,1}^2$

С другой стороны, поскольку в рамках линейной регрессии $\hat{\hat{x}}_0$ - линейная функция $\hat{x}_1$,

$$r_{\hat{x}_0,\hat{x}_i} = r_{\hat{\hat{x}}_0,\hat{x}_0}. \quad (6.22)$$

Из (6.21) и (6.22) следует, что коэффициент корреляции $r_{\hat{x}_0,\hat{x}_1}$ является мерой тесноты связи между случайной величиной $\hat{x}_0$ и случайной величиной $\hat{\hat{x}}_0$, определенной из линейной среднеквадратической регрессии (6.13). именно в этом смысле следует понимать утверждение, что коэффициент корреляции характеризует тесноту линейной связи между случайными величинами.

6.2.3 Двухкомпонентный анализ

Перейдем к рассмотрению линейной среднеквадратической регрессии связывающей $\hat{x}_0$ с двумя величинами $\hat{x}_1$ и $\hat{x}_2$:

$$\hat{\tilde{x}}_0 = c_0 + c_1\tilde{x}_1 + c_2\tilde{x}_2. \quad (6.23)$$

Минимизируя последовательность по c_0, c_1 и c_2 величину

$$M\left[\hat{\varepsilon}_{012}\right] = M\left[\left(\hat{\tilde{x}}_0 - \hat{x}_0\right)^2\right],$$

получим для определения коэффициента регрессии (6.23) систему уравнений:

$$\left.\begin{aligned} M[c_0 + c_1\hat{x}_1 + c_2\hat{x}_2 - \hat{x}_0] &= 0, \\ M[\hat{x}_1(c_0 + c_1\hat{x}_1 + c_2\hat{x}_2 - \hat{x}_0)] &= 0, \\ M[\hat{x}_2(c_0 + c_1\hat{x}_1 + c_2\hat{x}_2 - \hat{x}_0)] &= 0. \end{aligned}\right\} \quad (6.24)$$

Из первого уравнения системы (6.24) следует

$$c_0 = c_1 M_{\hat{x}_1} + c_2 M_{\hat{x}_2} - M_{\hat{x}_0}, \quad (6.25)$$

последние два уравнения (6.24) приводятся к виду:

$$\left.\begin{aligned} M[\hat{x}_1'(c_1\hat{x}_1' + c_2\hat{x}_2' - \hat{x}_0')] &= 0, \\ M[\hat{x}_2'(c_1\hat{x}_1' + c_2\hat{x}_2' - \hat{x}_0')] &= 0, \end{aligned}\right\}$$

откуда

$$\left.\begin{aligned} M[\hat{x}_0', \hat{x}_1'] &= c_1 M[\hat{x}_1'^2] + c_2 M[\hat{x}_1', \hat{x}_2'], \\ M[\hat{x}_0', \hat{x}_2'] &= c_1 M[\hat{x}_1', \hat{x}_2'] + c_2 M[\hat{x}_2'^2], \end{aligned}\right\} \quad (6.26)$$

где, как и ранее, $\hat{x}_k' = \hat{x}_k - M[\hat{x}_k]$.

Введя для упрощения записи обозначения

$$\delta_{\hat{x}_k} = \delta_k, \quad r_{\hat{x}_k,\hat{x}_1} = r_{k,1},$$

перепишем уравнения (6.26) в виде:

$$\left.\begin{aligned} \delta_0\delta_{0,1} &= c_1\delta_1 + c_2\delta_2 r_{1,2}, \\ \delta_0\delta_{0,2} &= c_1\delta_1 r_{1,2} + c_2\delta_2. \end{aligned}\right\} \quad (6.27)$$

Решения системы (6.27):

$$\left.\begin{aligned} c_1 &= \frac{\delta_0}{\delta_1}\,\frac{r_{0,1} - r_{0,2}r_{1,2}}{1 - r_{1,2}^2} \\ c_2 &= \frac{\delta_0}{\delta_2}\,\frac{r_{0,2} - r_{0,1}r_{1,2}}{1 - r_{1,2}^2} \end{aligned}\right\} \quad (6.28)$$

и формула (6.25) определяют регрессию (6.28).

Решение может быть записано более компактно, если ввести понятия частных дисперсий, частных средних квадратических отклонений коэффициентов корреляции.

В общем случае под частной дисперсией случайной величины $\hat{x}$ понимают дисперсию при фиксированных значениях других рассматриваемых величин. Аналогично определяются частное среднее квадратическое отклонение и частный коэффициент корреляции. Однако совместном нормальном распределении переменных указанные характеристики не зависят от значений, принятых фиксируемыми случайными величинами. В этом случае частная дисперсия $D_{\hat{x}_0,\hat{x}_1} =$

D_{0-1} определяется как «обычная» дисперсия ошибки линейной среднеквадратической регрессии $\hat{x}_0$ на $\hat{x}_1$:

$D_{0-1} = D[\hat{\hat{x}}_0 - \hat{x}_0]$,

где $\hat{\hat{x}}_0 = a_0 + a_1\hat{x}_1$.

Но из сравнения (6.25) с (6.23) следует, что $M[\hat{\hat{x}}_0] = M[\hat{x}_0]$, поэтому $D[\hat{\hat{x}}_0 - \hat{x}_0] = M\{(\hat{\hat{x}}_0 - \hat{x}_0 - M[\hat{\hat{x}}_0 - \hat{x}_0])^2\} = M[(\hat{\hat{x}}_0 - \hat{x}_0)^2] = M[\hat{\varepsilon}_{01}^2]$, и с учетом (6.20)

$$D_{0-1} = D_0(1 - r_{0,1}^2). \quad (6.29)$$

Аналогично определяются частные дисперсии $D_{\hat{x}_0,\hat{x}_2} = D_{0-2}$ и $D_{\hat{x}_1,\hat{x}_2} = D_{12}$:

$$D_{0-2} = D_0(1 - r_{0,2}^2). \quad (6.30)$$

$$D_{1-2} = D_1(1 - r_{1,2}^2). \quad (6.31)$$

Частные средние квадратические отклонения связаны с соответствующими частными дисперсиями формулой

$$\delta_{k,1} = \sqrt{D_{k,1}} \quad (6.32)$$

Частный коэффициент корреляции $r_{\hat{x}_0,\hat{x}_1,\hat{x}_2} = r_{0,1,2}$ при сделанном предположении определяется как «обычный» (<u>парный</u>) коэффициент корреляции между ошибками линейных среднеквадратических регрессий $\hat{x}_0$ на $\hat{x}_2$ и $\hat{x}_1$ на $\hat{x}_2$:

$$r_{0,1-2} = r_{(\hat{\hat{x}}_0 - \hat{x}_0)(\hat{\hat{x}}_1 - \hat{x}_1)}, \quad (6.33)$$

где $\hat{\hat{x}}_0 = b_0 + b_2\hat{x}_2$, $\hat{\hat{x}}_1 = d_0 + d_2\hat{x}_2$.

Учтем, что

$\hat{\hat{x}}_0 - \hat{x}_0 = \hat{\hat{x}}'_0 - \hat{x}'_0 = b_2\hat{x}'_2 - \hat{x}'_0$,

$\hat{\hat{x}}_1 - \hat{x}_1 = \hat{\hat{x}}'_1 - \hat{x}'_1 = d_2\hat{x}'_2 - \hat{x}'_1$,

и, воспользовавшись формулой (6.18) положив в ней последовательно $c_1 = b_2$ и $c_1 = d_2$, получим:

$$\left.\begin{aligned} \hat{\hat{x}}_0 - \hat{x}_0 &= \frac{\delta_0}{\delta_2} r_{0,2}\hat{x}'_2 - \hat{x}'_0, \\ \hat{\hat{x}}_1 - \hat{x}_0 &= \frac{\delta_1}{\delta_2} r_{1,2}\hat{x}'_2 - \hat{x}'_1. \end{aligned}\right\} \quad (6.34)$$

Введем частный корреляционный момент $k_{\hat{x}_0,\hat{x}_1,\hat{x}_2} = k_{0,1-2}$:

$k_{0,1-2} = k_{(\hat{\hat{x}}_0 - \hat{x}_0),(\hat{\hat{x}}_1 - \hat{x}_1)} = M[(\hat{\hat{x}}_0 - \hat{x}_0)'(\hat{\hat{x}}_1 - \hat{x}_1)']$.

Но, как уже отмечалось $\hat{\hat{x}}_0 - \hat{x}_0 = (\hat{\hat{x}}_0 - \hat{x}_0)'$, $\hat{\hat{x}}_1 - \hat{x}_1 = (\hat{\hat{x}}_1 - \hat{x}_1)'$, и согласно (6.34), после небольших преобразований найдем: $k_{0,1-2} = \delta_0\delta_1(r_{0,1} - r_{0,2}r_{1,2})$.

Из формул (6.30) и (6.31) $\delta_0^2 = D_{0-2}\frac{1}{1-r_{0,2}^2}$, $\delta_1^2 = D_{1-2}\frac{1}{1-r_{1,2}^2}$, и, таким образом, $k_{0,1-2} = \delta_{0-2}\delta_{1-2}\frac{r_{0,1}-r_{0,2}r_{1,2}}{\sqrt{(1-r_{0,2}^2)(1-r_{1,2}^2)}}$.

Следовательно частный коэффициент корреляции $r_{\hat{x}_0,\hat{x}_1,\hat{x}_2}$ может быть представлен в виде:

$$r_{0,1\,2} = \frac{k_{0,12}}{\delta_{0\,2}\delta_{1\,2}} = \frac{r_{0,1}-r_{0,2}r_{1,2}}{\sqrt{(1-r_{0,2}^2)(1-r_{1,2}^2)}}. \tag{6.35}$$

Аналогично

$$r_{0,2-1} = \frac{r_{0,1}-r_{0,2}1_{1,2}}{\sqrt{(1-r_{0,2}^2)(1-r_{1,2}^2)}}. \tag{6.36}$$

Формулы (6.35) и (6.36) дают связь частных коэффициентов корреляции с парными коэффициентами корреляции между соответствующими случайными величинами.

С учетом полученных выражений для частных средних квадратических отклонений и коэффициентов корреляции формулы для определения коэффициентов регрессии (6.28) приводятся к простому, легко запоминающемуся виду: *)

$$\left.\begin{aligned} c_1 &= \frac{\delta_{0-2}}{\delta_{1-2}}\, r_{0,1-2}, \\ c_2 &= \frac{\delta_{0-1}}{\delta}\, r_{0,2-1}. \end{aligned}\right\} \tag{6.37}$$

Перейдем к анализу качества регрессии (6.23). рассмотрим величину

$$\begin{aligned} &M[\hat{\varepsilon}_{012}^2] = M\left[\left(\hat{\tilde{x}}_0 - \hat{x}_0\right)^2\right] = M\left[\left(\hat{\tilde{x}}_0' - \hat{x}_0'\right)^2\right] = \\ &= M\left[\hat{\tilde{x}}_0'^2\right] + M[\hat{x}_0'^2] - 2\,M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right]. \end{aligned} \tag{6.38}$$

Ошибка регрессии $\left(\hat{\tilde{x}}_0 - \hat{x}_0\right)$, очевидно, не коррелирует с $\hat{\tilde{x}}_0$ - в противном случае рассчитанное значение $\hat{x}$ могло бы быть «подправлено» с учетом ошибки, что означало бы просто неоптимальность использованных коэффициентов регрессии.

Поэтому

$M\left[\hat{\tilde{x}}_0'\left(\hat{\tilde{x}}_0 - \hat{x}_0\right)'\right] = M\left[\hat{\tilde{x}}_0'^2\right] - M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right] = 0$

$M\left[\hat{\tilde{x}}_0'^2\right] = M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right]$.

Следовательно

$M[\hat{\varepsilon}_{012}^2] = M\left[\hat{\tilde{x}}_0'^2\right] - M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right] = D_0 - M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right]$.

Но, поскольку

$\hat{\tilde{x}}_0^1 = c_1\hat{\tilde{x}}_1^1 + c_2\hat{\tilde{x}}_2^1$,

$M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right] = c_1 M\left[\hat{\tilde{x}}_0', \hat{x}_1'\right] + c_2 M\left[\hat{\tilde{x}}_0', \hat{x}_2'\right]$

окончательно получим:

$M[\hat{\varepsilon}_{012}^2] = D_0 - \{c_1 M[\hat{x}_0', \hat{x}_1'] + c_2 M[\hat{x}_0', \hat{x}_2']\}$

или

$$M[\hat{\varepsilon}_{012}^2] = D_0 - (c_1 \delta_1 r_{0,1} + c_2 \delta_2 r_{0,2}) \delta_0. \quad (6.39)$$

Как видно, второе слагаемое первой части (6.39) равно сумме левых частей уравнений (6.26), использовавшихся для определения коэффициентов регрессии, помноженных соответственно на c_1 и c_2.

Преобразуем формулу (6.39) с учетом введенных понятий частных средних квадратических отклонений и коэффициентов корреляции.

В соответствии с первой формулой (6.28)

$$c_1 \delta_1 r_{0,1} = \delta_0 r_{0,1} \frac{r_{0,1} - r_{0,2} r_{1,2}}{1 - r_{1,2}^2}. \quad (6.40)$$

С другой стороны, из (6.36) следует:

$$r_{0,2} = r_{0,2\,1} \sqrt{(1 - r_{0,1}^2)(1 - r_{1,2}^2)} + r_{0,1} r_{1,2} \quad (6.41)$$

и согласно второй формуле (6.28)

$$c_2 \delta_2 r_{0,2} = \delta_2 \frac{\delta_0 \sqrt{(1 - r_{0,1}^2)}}{\delta_2 \sqrt{(1 - r_{1,2}^2)}} r_{0,2} r_{0,2-1}.$$

Поставим в правую часть полученной формулы выражение (6.41):

$$c_2 \delta_2 r_{0,2} = r_{0,2-1}^2 \delta_0 (1 - r_{0,1}^2) + r_{0,2-1} r_{0,1} r_{1,2} \frac{\delta_0 \sqrt{(1 - r_{0,1}^2)}}{\sqrt{(1 - r_{1,2}^2)}} = \\ = r_{0,2-1}^2 \delta_0 (1 - r_{0,1}^2) + \delta_0 r_{0,1} r_{1,2} \frac{r_{0,2} - r_{0,1} r_{1,2}}{1 - r_{1,2}^2}. \quad (6.42)$$

Просуммируем (6.40) и (6.42):

$$c_1 \delta_1 r_{0,1} + c_2 \delta_2 r_{0,2} = \delta_0 \left[r_{0,1} \frac{r_{0,1} - r_{0,2} r_{1,2}}{1 - r_{1,2}^2} + r_{0,2-1}^2 (1 - r_{0,1}^2) + r_{0,1} \frac{r_{0,2} r_{1,2} - r_{0,1} r_{1,2}^2}{1 - r_{1,2}^2} \right] = \delta_0 \left[r_{0,2-1}^2 (1 - r_{0,1}^2) + r_{0,1}^2 \right].$$

Таким образом, формула (6.39) приводится к виду

$$M[\hat{\varepsilon}_{012}^2] = D_0 - D_0 [r_{0,2-1}^2 (1 - r_{0,1}^2) + r_{0,1}^2] = D_0 (1 - r_{0,1}^2)(1 - r_{0,2-1}^2) \quad (6.43)$$

Аналогично может быть получено второе выражение:

$M[\hat{\varepsilon}_{012}^2] = D_0 (1 - r_{0,2}^2)(1 - r_{0,1-2}^2)$.

6.2.4 Многокомпонентный анализ

Коэффициенты множественной линейной среднеквадратической регрессии

$$\hat{\hat{x}}_0 = c_0 + c_1\hat{x}_1 + c_2\hat{x}_2 + \ldots + c_n\hat{x}_n \qquad (6.44)$$

удовлетворяют системе уравнений

$$\left.\begin{aligned} c_1\delta_1 + c_2\delta_2 r_{1,2} + \ldots + c_n\delta_n r_{1,n} &= \delta_0 r_{0,1}, \\ c_2\delta_2 r_{1,2} + c_2\delta_2 + \ldots + c_n\delta_n r_{2,n} &= \delta_0 r_{0,2}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . \\ c_2\delta_2 r_{1,n} + c_2\delta_2 r_{2n} + \ldots + c_n\delta_n &= \delta_0 r_{0,n}, \\ c_0 = M_{\hat{x}_0} - \left(c_1 M_{\hat{x}_1} + \ldots + c_n M_{\hat{x}_n}\right). \end{aligned}\right\} \qquad (6.45)$$

После нахождения коэффициентов $c_1, c_2, \ldots, c_n$ величина $M[\hat{\varepsilon}^2_{01\ldots n}]$, характеризующая качество регрессии, может быть определена по формуле

$$M[\hat{\varepsilon}^2_{01\ldots n}] = M\left[\left(\hat{\hat{x}}_0 - \hat{x}_0\right)^2\right] = D_0 - \delta_0 \sum_{i=1}^{n} c_i\delta_i r_{0,i}\,. \qquad (6.46)$$

С использованием понятий частных средних квадратических отклонений и коэффициентов корреляции решения системы (6.45) представляются в виде:

$$\left.\begin{aligned} c_1 &= r_{0,1-2,\ldots,n}\,\frac{\delta_{0-2,\ldots,n}}{\delta_{1-2,\ldots,n}}, \\ c_2 &= r_{0,2\ 1,3,\ldots,n}\,\frac{\delta_{0-1,3,\ldots,n}}{\delta_{2-1,3,\ldots,n}}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . \\ c_n &= r_{0,n-1,\ldots,n-1}\,\frac{\delta_{0-1,\ldots,n-1}}{\delta_{n-1,\ldots,n-1}}, \end{aligned}\right\} \qquad (6.47)$$

а в формуле (6.46) – в виде

$$M[\hat{\varepsilon}^2_{01\ldots n}] = D_0\left(1 - r^2_{0,1}\right)\left(1 - r^2_{0,2-1}\right) \ldots \left(1 - r^2_{0,n-1,\ldots n-1}\right). \qquad (6.48)$$

В формулах (6.47) и (6.48) частные коэффициенты корреляции имеют смысл парных коэффициентов корреляции между ошибками линейных среднеквадратических регрессий случайных величин, номера которых указаны в индексе перед точкой, на совокупность величин с номерами, указанными в индексе после точки. Так, например,

$$r_{0,2\ 1,3,4,\ldots,n} = r_{(\hat{\hat{x}}_0 - \hat{x}_0)(\hat{\hat{x}}_2 - \hat{x}_2)}\,, \qquad (6.49)$$

где $\hat{\hat{x}}_0 = a_0 + a_1\hat{x}_1 + a_3\hat{x}_3 + a_4\hat{x}_4 + \ldots + a_n\hat{x}_n$,

$\hat{\hat{x}}_0 = b_0 + b_1\hat{x}_1 + b_3\hat{x}_3 + b_4\hat{x}_4 + \ldots + b_n\hat{x}_n$.

Значения частных коэффициентов корреляции находятся последовательно по формулам вида:

$$\left.\begin{aligned} r_{0,1-2} &= \frac{r_{0,1} - r_{0,2}r_{1,2}}{\sqrt{\left(1 - r_{0,2}^2\right)\left(1 - r_{0,1-2}^2\right)}}, \\ r_{0,1-2,3} &= \frac{r_{0,1} - r_{0,2-3}r_{1,2-3}}{\sqrt{\left(1 - r_{0,2-3}^2\right)\left(1 - r_{0,1-2-3}^2\right)}}, \end{aligned}\right\} \qquad (6.50)$$

Частные средние квадратические отклонения в формулах (6.47) имеют смысл средней квадратической ошибки регрессии (линейной, среднеквадратической) случайной величины, номер которой указан, в индексе перед точкой совокупность величин с номерами, указанными в индексе после точки. Так, например,

$\delta_{0,2\ 1,3,4,\ldots,n} = \delta_{(\hat{\tilde{x}}_0 - \hat{x}_0)}$,

Значения частных средних квадратических отклонений могут быть определены последовательно по формулам вида:

$$\left.\begin{aligned} \delta_{0-1}^2 &= \delta_0^2\left(1 - r_{0,1}^2\right), \\ \delta_{0-1,2}^2 &= \delta_{0-1}^2\left(1 - r_{0,2-1}^2\right), \end{aligned}\right\} \qquad (6.51)$$

и т. д. *)

формуле (6.46) для анализа качества линейной среднеквадратической регрессии (6.44) можно придать более компактный вид, если ввести в рассмотрение множественный (сводный) коэффициент корреляции R, имеющий смысл парного коэффициента корреляции между случайными величинами $\hat{\tilde{x}}_0$ и $\hat{x}_0$:

$$R = r_{\hat{\tilde{x}}_0, \hat{x}_0}. \qquad (6.52)$$

Действительно, поскольку $M[\hat{\tilde{x}}_0] = M[\hat{x}_0]$,

$M[\hat{\varepsilon}_{01\ldots n}^2] = M\left[\left(\hat{\tilde{x}}_0 - \hat{x}_0\right)^2\right] = M\left[\left(\hat{\tilde{x}}_0' - \hat{x}_0'\right)^2\right] =$

$= M\left[(\hat{\tilde{x}}_0')^2\right] + M[(\hat{x}_0')^2] - 2\, M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right]$.

Но, как отмечалось выше, $M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right] = M\left[(\hat{\tilde{x}}_0')^2\right]$. Поэтому

$$M\left[\hat{\delta}_{01\ldots n}^2\right] = D_0\, M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right] = D_0\left(1 - \frac{M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right]}{D_0}\right) = D_0\left(1 - \frac{M\left[\hat{\tilde{x}}_0', \hat{x}_0'\right] D\left[\hat{\tilde{x}}_0\right]}{D_0 D\left[\hat{\tilde{x}}_0\right]}\right) =$$

$$= D_0\left(1 - \frac{M^2\left[\hat{\tilde{x}}_0', \hat{x}_0'\right]}{D_0[\hat{x}_0] D_0\left[\hat{\tilde{x}}_0\right]}\right) = D_0(1 - R^2).$$

Таким образом,

$$M[\hat{\varepsilon}_{01\ldots n}^2] = D_0(1 - R^2), \qquad (6.53)$$

и, с учетом (6.48),

$$R^2 = 1 - (1 - r_{0,1}^2)\left(1 - r_{0,2-1}^2\right) \ldots \left(1 - r_{0,n-1,2,\ldots n-1}^2\right). \qquad (6.54)$$

6.2.5 Выборочный регрессионный анализ

В предыдущих параграфах при рассмотрении регрессионного анализа предполагалось известным совместное распределение предиктанта и предикто-

ров. В практической работе это распределение, как правило, неизвестно и построение регрессионных моделей приходится основывать на результатах статистической обработки ограниченной выборки (архива), содержащей реализации случайного вектора $\hat{\chi}_{<n+1>}$. Регрессии, полученные на материале ограниченной выборки, принято называть выборочными или эмпирическими.

Основной задачей теории выборочного регрессионного анализа является нахождение правила оценивания регрессии

$\tilde{x}_0 = f(\hat{\chi}_{<n>})$,

Обеспечивающего понимаемую в вероятностном смысле максимальную близость значений $f(\hat{\chi}_{<n+1>})$ к значениям предиктанта $\hat{x}_0$. В метеорологической практике, как уже отмечалось, используется преимущественно среднеквадратическая регрессия, минимизирующая математическое ожидание квадрата ошибки регрессии (6.3), причем между предиктантом и предикторами предполагается линейной.

Уравнение выборочной множественной линейной регрессии, соответствующей «теоретической» регрессии (6.44), имеет вид:

$$\hat{\tilde{x}}_0 = c_0^* + c_1^* \hat{x}_1 + c_2^* \hat{x}_2 + \ldots + c_n^* \hat{x}_n, \qquad (6.55)$$

где $c_0^*, c_1^*, \ldots + c_n^*$ - выборочные оценки аналогичных коэффициентов регрессии (6.44).

оценивание коэффициентов регрессии производится по результатам наблюдений – реализациям случайного вектора $\hat{\chi}_{<n+1>}$ (табл. 6.1).

таблица 6.1

Номера реализаций	Значения предиктанта x_0 и предикторов $x_1, x_2, \ldots, x_n$					
	x_0	x_1	…	x_i	…	x_n
1	x_{01}	x_{11}	…	x_{i1}	…	x_{n1}
2	x_{02}	x_{12}	…	x_{i2}	…	x_{n2}
…	…	…	…	…	…	…
j	x_{0j}	x_{1j}	…	x_{ij}	…	x_{nj}
…	…	…	…	…	…	…
m	x_{0m}	x_{1m}	…	x_{im}	…	x_{nm}

Основным методом оценивания коэффициентов среднеквадратической регрессии является метод наименьших квадратов.

6.2.6 Метод наименьших квадратов

Если воспользоваться в качестве выборочной оценки математического ожидания случайной величины $\left(\hat{\tilde{x}}_0 - \hat{x}_0\right)^2$ средним арифметическим из ее m зна-

чений: $\frac{1}{m}\sum_{j=1}^{m}(\tilde{x}_{0j} - x_{0j})^2 = \overline{(\tilde{x}_0 - x_0)^2}$, то логично предположить что для выполнения требования (6.3) оценки коэффициентов регрессии должно удовлетворять условию *)

$$\overline{(\tilde{x}_0 - x_0)^2} = \frac{1}{m}\sum_{j=1}^{m}(c_0^* + c_1^* x_{1j} + c_2^* x_{2j} + \ldots + c_n^* x_{nj} - x_{cj})^2 = \min_{(c_0, c_1, \ldots, c_n)} \quad (6.56)$$

Условию (6.56) соответствует система уравнений

$$\frac{\partial}{\partial c_i}\overline{(\tilde{x}_0 - x_0)^2} = 0; \qquad i = 0,1, \ldots, n. \quad (6.57)$$

Метод оценивания коэффициентов регрессии, основанный на решении системы (6.57), называется методом наименьших квадратов, а полученные таким образом оценки – МНК-оценками.

Первое уравнение системы (при i=0) имеет вид:

$$\overline{x_0} = c_0^* + c_1^*\overline{x_1} + c_2^*\overline{x_2} + \ldots + c_n^*\overline{x_n}. \quad (6.58)$$

Отсюда, учтя (6.55), получаем $\overline{x_0} = \bar{\tilde{x}}_0$, и уравнения (6.57) могут быть для i = 1, 2, …, n, представлены в виде

$$\frac{\partial}{\partial c_i}\overline{(\tilde{x}_0' - x_0')^2} = 0; \qquad i = 1,2, \ldots, n, \quad (6.59)$$

где $\tilde{x}_0' = \tilde{x}_0 - \bar{\tilde{x}}_0$, $x_0' = x_0 - \overline{x_0}$.

Вводя аналогичные обозначения для всех предикторов, запишем систему (6.59) следующим образом:

$$\left.\begin{aligned}
\overline{x_0'x_1'} &= c_1^*\overline{(x_1')^2} + c_2^*\overline{x_1'x_2'} + \ldots + c_n^*\overline{x_1'x_n'},\\
\overline{x_0'x_2'} &= c_1\overline{x_1'x_2'} + c_2\overline{(x_2')^2} + \ldots + c_n\overline{x_2'x_n'},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
\overline{x_0'x_n'} &= c_1\overline{x_1'x_n'} + c_2\overline{x_2'x_n'} + \ldots + c_n\overline{(x_n')^2}.
\end{aligned}\right\} \quad (6.60)$$

Использовав введенные в подразд. 4.7 и 4.8 статистические оценки дисперсии, корреляционного момента и коэффициента корреляции

$$D_k^* = D_{\hat{x}_k}^* = \delta_k^{*2} = \delta_{\hat{x}_k}^{*2} = \frac{1}{m-1}\sum_{j=1}^{m}\left(x_{kj}'\right)^2,$$

$$K_{k,1}^* = K_{\hat{x}_k x_1} = \frac{1}{m-1}\sum_{j=1}^{m} x_{kj}'\, x_{1j}',$$

$$r_{k,1}^* = K_{\hat{x}_k\hat{x}_1} = \frac{K_{\hat{x}_k\hat{x}_1}}{\delta_{\hat{x}_k}^*\delta_{\hat{x}_1}^*}.$$

уравнение (6.60) нетрудно преобразовать к виду:

$$\left.\begin{aligned}
\delta_0^* r_{0,1}^* &= c_1^*\delta_1^* + c_2^*\delta_2^* r_{1,2}^* + \ldots + c_n^*\delta_n^* r_{1,n}^*,\\
\delta_0^* r_{0,2}^* &= c_1^*\delta_1^* r_{1,2}^* + c_2^*\delta_2^* + \ldots + c_n^*\delta_n^* r_{2,n}^*,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
\delta_0^* r_{0,n}^* &= c_1^*\delta_1^* r_{1,n}^* + c_2^*\delta_2^* r_{2,n}^* + \ldots + c_n^*\delta_n^*.
\end{aligned}\right\} \quad (6.61)$$

Сравнение уравнений (6.58) и (6.61) с уравнениями (6.45) показывает, что все формулы подраз. 6.2.4 выполняются и для выборочной регрессии с заменой содержащихся в них числовых характеристик генерального распределения соответствующими статистическими оценками.

Для иллюстрации методов выборочного регрессионного анализа ниже будет рассмотрен ряд примеров, связанных с задачей регрессионного прогноза максимальной температуры воздуха в дневные часы на ст. Омск по результатам утреннего зондирования атмосферы на той же станции. Значения предиктанта x(℃) и предикторов x_1, x_2, …, x_8 приводятся в табл. 6.2. Здесь x_1 и x_2 - температура воздуха у поверхности земли и на поверхности 850 гПа соответственно (℃), x_3 - дефицит точки росы на поверхности 850 гПа (℃),x_4 - количество облачности нижнего яруса (баллы), x_5 - скорость ветра у поверхности земли ($мс^{-1}$), x_6 и x_7 - количество облачности среднего и верхнего ярусов (баллы) и x_8 - относительная влажность воздуха у поверхности земли (%).

Таблица 6.2

Значения предикторов x_1, x_2, …, x_8 и предиктанта x_0

Номер варианта	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_0
1	10	8	9	0	9	0	0	60	20
2	11	7	6	0	7	0	3	65	19
3	11	5	1	10	5	10	10	90	13
4	7	3	2	4	5	0	0	90	13
5	14	11	4	4	10	4	8	92	21
6	14	8	1	10	7	0	0	94	18
7	8	5	3	7	4	0	0	88	13
8	9	12	7	2	5	0	0	80	26
9	15	11	1	9	8	10	0	90	21
10	16	10	3	5	12	5	8	74	24
11	12	6	1	10	10	10	10	80	15
12	10	13	6	0	8	0	10	82	24
13	17	9	8	2	7	3	3	85	30
14	12	7	4	4	5	5	5	70	20
15	10	11	3	5	3	7	7	75	21
16	9	10	3	8	10	10	10	94	14
17	16	15	7	2	2	5	5	86	27
18	14	12	10	0	2	0	0	75	23
19	13	12	7	0	2	5	10	77	24

20	9	6	5	1	1	0	0	75	24
21	17	18	3	4	4	0	0	72	29
22	19	16	2	8	3	8	8	84	24
23	17	13	11	0	7	0	4	68	32
24	15	17	12	0	6	2	6	70	31
25	17	14	6	0	4	4	7	65	29
26	20	15	4	2	3	6	6	70	29
27	12	11	2	8	2	4	9	80	18
28	11	8	5	1	8	5	6	90	19
29	9	10	8	0	5	3	5	69	18
30	6	3	8	0	4	0	2	74	13
31	4	4	3	4	7	7	10	87	14
32	3	4	6	0	3	0	4	85	13
33	2	6	1	10	7	10	10	65	7
34	1	4	2	8	5	5	7	74	12
35	4	6	9	0	7	0	4	86	17
36	5	4	5	2	5	2	5	90	13
37	3	0	10	0	4	5	5	96	12
38	2	-2	5	3	6	5	7	97	11
39	1	0	3	6	4	10	10	92	12
40	4	1	4	3	1	10	10	90	13
Выборка 2									
41	5	7	5	1	5	3	5	70	16
42	12	14	4	3	10	2	7	95	25
43	6	4	3	3	9	7	10	85	11
44	3	6	5	3	7	4	4	78	17
45	13	9	8	0	7	9	10	85	19
46	2	-1	2	6	12	10	10	95	14
47	14	10	11	0	10	4	7	60	25
48	10	16	2	10	4	10	10	90	17
49	18	15	8	0	4	4	6	82	28
50	18	16	5	0	5	4	7	85	30
51	18	14	4	3	5	3	4	86	24
52	16	18	7	0	5	0	0	74	30
53	14	12	3	5	5	7	10	78	18
54	8	4	9	0	6	0	4	84	15
5	5	3	0	9	0	5	5	79	13

56	2	3	5	2	3	4	4	72	14
57	5	6	4	2	10	4	5	79	16
58	6	3	7	0	7	4	8	94	11
59	2	-1	4	3	3	10	10	95	10
60	7	5	4	2	2	7	7	90	19

При выполнении расчетов выборка 1 («обучающая») будет использоваться непосредственно для оценивания коэффициентов регрессии, а выборка 2 («экзаменационная») – для проверки качества оценивания.

В табл. 6.3 и 6.4 приводятся полученные по первой выборке и округленные до третьего знака после запятой оценки математического ожидания, дисперсии и среднего квадратического отклонения для рассматриваемых случайных величин и парных коэффициентов корреляции.

Таблица 6.3

Оценки $\mathrm{M}[\hat{\mathrm{x}}_i]$, $D[\hat{\mathrm{x}}_i]$ и $\delta[\hat{\mathrm{x}}_i]$

Выборка 1

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_0
$\mathrm{M}^*_{\hat{\mathrm{x}}_i}$	10.225	8.325	4.975	3.550	5.425	4.000	5.350	80.650	19.375
$D^*_{\hat{\mathrm{x}}_i}$	28.794	24.328	9.307	12.510	7.174	13.641	13.515	101.521	43.010
$\delta^*_{\hat{\mathrm{x}}_i}$	5.366	4.932	3.051	3.537	2.678	3.693	3.676	10.076	6.558

Таблица 6.4

Оценки парных коэффициентов корреляции

Выборка 1

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_0
x_1	1	0,837	0,079	-0,026	0,077	-0,085	-0,163	-0,306	0,838
x_2	0,837	1	0,174	-0,106	0,005	-0,149	-0,087	-0,436	0,840
x_3	0,079	0,174	1	-0,857	-0,102	-0,521	-0,262	-0,291	0,417
x_4	-0,026	-0,106	-0,857	1	0,207	0,597	0,265	0,294	0,384
x_5	0,077	0,005	-0,102	0,207	1	0,093	0,125	0,039	-0,046
x_6	-0,085	-0,149	-0,521	0,597	0,093	1	0,699	0,263	-0,307
x_7	-0,163	-0,087	-0,262	0,265	0,125	0,699	1	0,136	-0,233
x_8	-0,306	-0,436	-0,291	0,294	0,039	0,263	0,136	1	-0,414
x_0	0,838	0,840	0,417	-0,384	-0,046	-0,307	-0,233	-0,414	1

В качестве первого примера рассмотрим задачу разработки регрессионных прогнозов максимальной дневной температуры воздуха у поверхности земли на ст. Омск (x_0) по дефициту точки росы на поверхности 850 гПа (x_2) и количеству

облачности нижнего яруса (x_4) в утренние часы на той же станции.

Пример 1. Оценить по выборке 1 (табл. 6.1) значения коэффициентов и ошибки выборочной регрессии

$$\tilde{x}_0 = c_0^* + c_2^* \hat{x}_2 + c_4^* \hat{x}_4.$$

В соответствии с формулами подразд. 6.2.3

$$c_2^* = \frac{\delta_{0-4}^*}{\delta_{1-4}^*} r_{0,2-4}^*,$$

$$c_4^* = \frac{\delta_{0-2}^*}{\delta_{4-2}^*} r_{0,2-4}^*,$$

$$c_0^* = \overline{x_0} - c_2^* \overline{x_2} - c_4^* \overline{x_4},$$

$$r_{0,2-4}^* = \frac{r_{0,2}^* - r_{0,4}^* r_{2,4}^*}{\sqrt{(1 - r_{0,4}^{*2})(1 - r_{2,4}^{*2})}},$$

$$r_{0,4-2}^* = \frac{r_{0,4}^* - r_{0,2}^* r_{2,4}^*}{\sqrt{(1 - r_{0,2}^{*2})(1 - r_{2,4}^{*2})}},$$

$$\delta_{0-4}^* = \delta_0^* \sqrt{1 - r_{0,4}^{*2}},$$

$$\delta_{0-2}^* = \delta_0^* \sqrt{1 - r_{0,2}^{*2}},$$

$$\delta_{2-4}^* = \delta_2^* \sqrt{1 - r_{2,4}^{*2}},$$

$$\delta_{4-2}^* = \delta_4^* \sqrt{1 - r_{4,2}^{*2}},$$

$$\overline{(\tilde{x}_0 - x_0)^2} = D_0^* \left(1 - r_{0,2}^*\right)\left(1 - r - r_{0,4-2}^*\right).$$

Воспользовавшись оценками, приведенными в табл. 6.3 и 6.4, 6.2.4 найдем: δ_{0-4}^* = 6,055, δ_{0-2}^* = 3,558, δ_{2-4}^* = 4,905, δ_{4-2}^* = 3,517; $r_{0,2-4}$ = 0,871, $r_{0,4-2}$ = - 0,547; c_2 = 1,075, c_4 = - 0,553, c_0 = 12,389, и, следовательно

$$\hat{\tilde{x}}_0 = 12{,}389 + 1{,}075\hat{x}_2 - 0{,}553\hat{x}_4,$$

причем $\overline{(\tilde{x}_0 - x_0)^2}$ = 8,878.

Этим примером мы закончим рассмотрение вопросов нахождения МНК-оценок коэффициентов регрессии и перейдем к анализу качества оценивания.

Основными задачами анализа качества оценивания параметров и ошибок регрессии являются определение точности и надежности соответствующих выборочных оценок. Рассмотрение этих задач начнем с простейшего случая, когда статистическая зависимость предиктанта $\hat{x}_0$ от предиктора $\hat{x}_1$ задается линейным уравнением

$$x_0 = c_0 + c_1 x_1 + \varepsilon. \qquad (6.62)$$

МНК-оценки коэффициентов с и с можно в соответствии с формулами (6.58) и (6.59) представить в виде:

$$c_0^* = \bar{x}_0 - c_1^* \bar{x}_1 = \frac{1}{m}\sum_{j=1}^{m} x_{0j} - c_1^* \bar{x}_1, \tag{6.63}$$

$$c_1^* = \frac{x_0' x_1'}{(x_1')^2} = \frac{\sum_{j=1}^{m} x_{0j}' x_{1j}'}{\sum_{j=1}^{m} (x_{1j}')^2} = \frac{\sum_{j=1}^{m} x_{0j} x_{1j}'}{\sum_{j=1}^{m} (x_{1j}')^2}. \tag{6.64}$$

При анализе качества оценивания коэффициента c_0 и c_1 выборочные оценки этих коэффициентов должны для каждого значения предиктора считаться линейными величинами. Случайными являются и ошибки регрессии для этих значений предиктора. Покажем, что если ошибка $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots$ независимыми и имеют нулевое математическое ожидание и одинаковую дисперсию $\delta_{\hat{\varepsilon}}^2$, то оценки (6.63) и (6.64) удовлетворяют требованиям несмещенности и состоятельности. Действительно, поскольку для любого фиксированного x_1 по предположению $M[\hat{\varepsilon}] = 0$ и в силу (6.62), $M[\hat{x}_{0j}] = c_0 + c_1 x_{1j}$,

$$M[\hat{c}_1^*] = \frac{\sum_{j=1}^{m} x_{1j}' M[\hat{x}_{0j}]}{\sum_{j=1}^{m} (x_{1j}')^2} = \frac{\sum_{j=1}^{m} x_{1j}' (c_0 + c_1 x_{1j})}{\sum_{j=1}^{m} (x_{1j}')^2} =$$

$$= c_1 \frac{\sum_{j=1}^{m} x_{1j}' x_{1j}}{\sum_{j=1}^{m} (x_{1j}')^2} = c_1,$$

$$M[\hat{c}_0^*] = \frac{1}{m}\sum_{j=1}^{m} M[\hat{x}_{0j}] - c_1^* \overline{x_1} = \frac{1}{m}\sum_{j=1} (c_0 + c_1 x_{1j}) - c_1^* \overline{x_1} = c_0,$$

Таким образом, обе оценки являются несмещенными.

С другой стороны, поскольку при фиксированном значении $x_1 = x_{1j}$ в силу (6.62) $D_{\hat{x}_{0j}} = D_{\hat{\varepsilon}_j}$ и по предположению $D_{\hat{\varepsilon}_j} = \delta_{\hat{\varepsilon}}^* = \text{const}$, из (6.64) получаем

$$D[\hat{c}_1^*] = \frac{D[\sum_{j=1}^{m} \hat{x}_{0j} x_{1j}']}{\left\{\sum_{j=1}^{m} (x_{1j}')^2\right\}^2} = \frac{\sum_{j=1}^{m} (x_{1j}'^2)}{\left\{\sum_{j=1}^{m} (x_{1j}')^2\right\}^2} D_{\hat{x}_0} \frac{\delta_{\hat{\varepsilon}}^2}{\sum_{j=1}^{m} (x_{1j}')^2}, \tag{6.65}$$

а из (6.63), учитывая некоррелированность $\hat{x}_{0j}$ и $\hat{c}_1^*$,

$$D[\hat{c}_0^*] = D\left[\frac{1}{m}\sum_{j=1}^{m} \hat{x}_{0j}\right] + D[\overline{x_1}\, \hat{c}_1^*] = \frac{1}{m} D_{\hat{x}_0} - \bar{x}_1^2\, D[\hat{c}_1^*]$$

или, воспользовавшись (6.65),

$$D[\hat{c}_0^*] = \left\{\frac{1}{m} + \frac{\bar{x}_1^2}{\sum_{j=1}^{m} (x_{1j}')^2}\right\} \delta_{\hat{\varepsilon}}^2, \tag{6.66}$$

Поскольку при $m \to \infty$ оба полученные выражения стремятся к нулю, рассматриваемые МНК-оценки c_0 и c_1 состоятельны.

Для проведения дальнейшего анализа предположим, что при каждом фиксированном значении предиктора $x_{11}, x_{12}, \ldots$ ошибки регрессии $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots$ распределены нормально. При этом условии в силу (6.62) нормально распределены и $\hat{x}_{01}, \hat{x}_{02}, \ldots$. Но, согласно (6.64), $\hat{c}_1^*$ - линейная функция $x_{01}, x_{02}, \ldots$, и, следова-

тельно, случайная величина $\hat{c}_1^*$ также распределена нормально (с математическим ожиданием с и дисперсией $\frac{\delta_{\hat{\varepsilon}}^2}{\sum_{j=1}^m \left(x'_{1j}\right)^2}$).

Не трудно убедиться, что оценка дисперсии $\hat{\varepsilon}\delta_{\hat{\varepsilon}}^2 = \frac{1}{m}\sum_{j=1}^m \varepsilon_j^2$ смещена относительно $\delta_{\hat{\varepsilon}}^{*2}$, а несмещенной является оценка

$$\delta_{\hat{\varepsilon}}^{*2} = \frac{1}{m-2}\sum_{j=1}^m \varepsilon_j^2 = \frac{1}{m-2}\sum_{j=1}^m \left(x_{0j} - c_0^* - c_1^* x_{1j}\right)^2. \quad (6.67)$$

Отношение

$$(m-2)\frac{(\hat{\delta}_{\hat{\varepsilon}}^*)^2}{\delta_{\hat{\varepsilon}}^2} = \sum_{j=1}^m \frac{\hat{\varepsilon}_j^2}{\delta_{\hat{\varepsilon}}^2}, \quad (6.68)$$

при нормальном распределении $\hat{\varepsilon}_1$, $\hat{\varepsilon}_2$, … с одинаковой дисперсией $\delta_{\hat{\varepsilon}}^2$ и нулевым математическим ожиданием имеет распределение χ^2 с $(m-2)$ степенями свободы. Следовательно $100(1-\beta)$%-ный доверительный интервал для $\delta_{\hat{\varepsilon}}^2$ равен

$$\frac{m-2}{\chi^2_{m-2,1-\beta/2}}\delta_{\hat{\varepsilon}}^{*2} < \delta_{\hat{\varepsilon}}^2 < \frac{m-2}{\chi^2_{m-2,1-\beta/2}}\delta_{\hat{\varepsilon}}^{*2}. \quad (6.69)$$

Поскольку, далее, случайные величины $\frac{\chi_{m-2}}{m-2}$ и $\hat{c}_1^*$ независимы, отношение $\frac{\hat{c}_1^* - c_1}{\sqrt{\frac{\sum_{j=1}^m \left(x'_{1j}\right)^2}{(\delta_{\hat{\varepsilon}}^*)^2}}}$.

подчинено распределению Стьюдента с $(m-2)$ степенями свободы .

Поэтому $100(1-\beta)$%-ный доверительный интервал для c_1 принимает вид:

$$c_1^* - \frac{t_{m-2,1-\beta/2}}{\sqrt{\sum_{j=1}^m \left(x_{1j} - \overline{x_1}\right)^2}}\delta_{\hat{\varepsilon}}^* < c_1 < c_1^* + \frac{t_{m-2,1-\beta/2}}{\sqrt{\sum_{j=1}^m \left(x_{1j} - \overline{x_1}\right)^2}}\delta_{\hat{\varepsilon}}^*, \quad (6.70)$$

а для c_0:

$$c_0^* - t_{m-2,1-\beta/2}\left[\frac{1}{m} + \frac{\bar{x}_1^2}{\sum_{j=1}^m \left(x_{1j} - \overline{x_1}\right)^2}\right]^{1/2}\delta_{\hat{\varepsilon}}^* < c_0 < c_0^* + t_{m-2,1-\beta/2}\left[\frac{1}{m} + \frac{\bar{x}_1^2}{\sum_{j=1}^m \left(x_{1j} - \overline{x_1}\right)^2}\right]^{1/2}\delta_{\hat{\varepsilon}}^* \quad (6.71)$$

Перейдем к анализу качества оценивания предиктанта. Сделаем это вначале для фиксированного значения предиктора x_1^0. Выборочная регрессия, соответствующая (6.62), имеет вид:

$\tilde{x}_0^* = c_0^* + c_1^* x_1$.

Так как оценка $\tilde{x}_0^*$ не смещена относительно $\overline{x_0}$,

$c_0^* = \overline{x_0} - c_1^* \overline{x_1}$.

При $\hat{x}_1 = x_1^0$

$\tilde{x}_0^*[\hat{x}_1 = x_1^0] = c_0^* + c_1^* x_1^0 = x_0 + c_1^*(x_1^0 - x_1)$,

откуда

$D[\hat{x}_0^*/\hat{x}_1 = x_1^0] = D[\hat{x}_0] + (x_1^0 - \overline{x_1})^2 D[\hat{c}_1^*]^2.$

Но

$$D[\bar{\hat{x}}_0] = D\left[\frac{1}{m}\sum_{j=1}^{m}\hat{x}_{0j}\right] = \frac{1}{m^2}mD_{\hat{x}_0} = \frac{\delta_{\hat{\varepsilon}}^2}{m},$$

и, с учетом (6.65)

$$D[\hat{x}_0^*/\hat{x}_1 - x_1^0] - \left\{\frac{1}{m} + \frac{(x_1^0-\overline{x_1})^2}{\sum_{j=1}^{m}(x_{1j}-\overline{x_1})^2}\right\}\delta_{\hat{\varepsilon}}^2. \quad (6.72)$$

Вводя оценку (6.67) дисперсии ошибки и воспользовавшись при построении доверительного интервала распределением Стьюдента, получим для 100(1 – β)%-ного интервала:

$$x_{0_{\hat{x}_1 = x_1^0}}^* - t_{m-2,1-\beta/2}\left[\frac{1}{m} + \frac{(x_1^0-\overline{x_1})^2}{\sum_{j=1}^{m}(x_{1j}-\overline{x_1})^2}\right]^{1/2}\delta_{\hat{\varepsilon}}^{*2} <$$

$$< x_{0_{\hat{x}_1 = x_1^0}} < \quad (6.73)$$

$$< x_{0_{\hat{x}_1 = x_1^0}}^* + t_{m-2,1-\beta/2}\left[\frac{1}{m} + \frac{(x_1^0-\overline{x_1})^2}{\sum_{j=1}^{m}(x_{1j}-\overline{x_1})^2}\right]^{1/2}\delta_{\hat{\varepsilon}}^{*2}.$$

Формула (6.73) определяет интервал, в котором с заданной доверительной вероятностью находится значение предиктанта, соответствующее конкретному выборочному значению предиктора x_1^0. Вообще говоря, разным x_1^0 соответствуют различные доверительные интервалы. Для анализа качества регрессионной модели наибольший интерес представляет определение доверительной области для всей линии регрессии, то есть области, в которой должна с заданной доверительной вероятностью находится «истинная» линия регрессии.

Можно показать [3], что с вероятностью $1 - \beta$ значения предиктанта при любых x лежат в интервале

$$x_{0_{\hat{x}_1 = x_0^0}}^* \pm d\delta_{\hat{\varepsilon}}^*\left[\frac{1}{m} + \frac{(x_1^0-\overline{x_1})^2}{\sum_{j=1}^{m}(x_{1j}-\overline{x_1})^2}\right]^{1/2}, \quad (6.74)$$

где $d = \sqrt{2}\sqrt{F_{2,m-2,1-\beta}}$, $F_{2,m-2,1-\beta}$ - значение случайной величины $\hat{F}_{2,m-2}$, распределенной по закону Фишера, соответствующее вероятности $1 - \beta$. В табл. 4 Приложения приводятся значения $F_{2,m-2}$ для $1 - \beta = 0{,}05$ (верхние полустровки) и $1 - \beta = 0{,}01$ (нижние полустровки).

Пример 2. По данным табл. 6.3 и 6.4 оценить ошибку коэффициенты линейной среднеквадратической регрессии $\hat{x}_0$ на $\hat{x}_2$ и проанализировать качество оценивания.

Оцениваются коэффициенты регрессии

$\hat{x}_0 = c_0 + c_2\hat{x}_2 + \hat{\varepsilon}.$

В соответствии с формулами 6.2.3

$c_2^* = r^*_{\hat{x}_0\hat{x}_2}\frac{\delta_{\hat{x}_0}}{\delta^*_{\hat{x}_2}} = 0{,}840\,\frac{6{,}558}{4{,}932} = 1{,}117,$

$c_0^* = \overline{x_0} - c_2^*\overline{x_2} = 19{,}375 - 1{,}117 - 8{,}325 = 10{,}076.$

Следовательно, выборочная регрессия $\hat{x}_0$ на $\hat{x}_2$:

$\tilde{x}_0 = 10{,}076 + 1{,}117\,x_2.$

При этом $\bar{\varepsilon} = 0$, а несмещенная оценка дисперсии $\hat{\varepsilon}$ по (6.67) равна

$\delta^{*2}_{\hat{\varepsilon}} = \frac{1}{m-2}\sum_{j=1}^{m}\varepsilon_j^2 = \frac{m-1}{m-2}\left\{\frac{1}{m-1}\sum_{j=1}^{m}\varepsilon_j^2\right\} = = \frac{m-1}{m-2}D^*_{\hat{x}_0}(1 - r^{*2}_{\hat{x}_0\hat{x}_2}) =$

$= \frac{39}{38}\,43{,}010(1 - 0{,}840^2) = 12{,}995.$

Переходим к анализу качества оценивания коэффициентов регрессии. Вновь обратившись к табл. 6.3 и 6.4, найдем:

$\sum_{j=1}^{m}\left(x_{2j} - \overline{x_2}\right)^2 = (m-1)\,D^*_{\hat{x}_0} = 39 - 24{,}328 = 948{,}792,$

$\frac{1}{m} + \frac{\bar{x}_2^2}{\sum_{j=1}^{m}\left(x_{2j} - \overline{x_2}\right)^2} = \frac{4}{40} + \frac{8{,}325^2}{948{,}792} = 0{,}098,$

$\frac{\delta^*_{\hat{\varepsilon}}}{\sqrt{\sum_{j=1}^{m}\left(x_{2j} - \overline{x_2}\right)^2}} = \sqrt{\frac{12{,}995}{948{,}792}} = 0{,}117,$

$\delta^*_{\hat{\varepsilon}}\sqrt{\frac{1}{m} + \frac{\bar{x}_2^2}{\sum_{j=1}^{m}\left(x_{2j} - \overline{x_2}\right)^2}} = \sqrt{12{,}995 - 0{,}098} = 1{,}128.$

Для доверительной вероятности 0,99 по табл. 2 (Приложение) находим $t_{38;0{,}01} = 2{,}796$, и в соответствии с (6.70) и (6.71), доверительные границы равны:

для c_0 $10{,}076 \pm 1{,}398 - 1{,}128 = 10{,}076 \pm 1{,}577,$

для c_2 $1{,}117 \pm 1{,}398 - 0{,}117 = 1{,}117 \pm 0{,}164.$

Определим, наконец, границы доверительной области для линии регрессии. Для той же доверительной вероятности 0,99 по табл. 4 Приложения $F_{2,38;001}$, и с учетом полученных выше значений $\overline{x_2}$, $\sum_{j=1}^{m}\left(x_{2j} - \overline{x_2}\right)^2$ и $\delta^*_{\hat{\varepsilon}}$ формула (6.74) принимает вид:

$$x^*_{0_{\hat{x}_2 = x_2^0}} \pm \sqrt{2\,5{,}21\left[\frac{1}{40} + \frac{\left(x_2^0 - 8{,}325\right)^2}{948\ 792}\right]}\ 12{,}995.$$

По результатам расчетов для различных значений x_2 построены границы доверительной области, показанные на рис. 6.1.

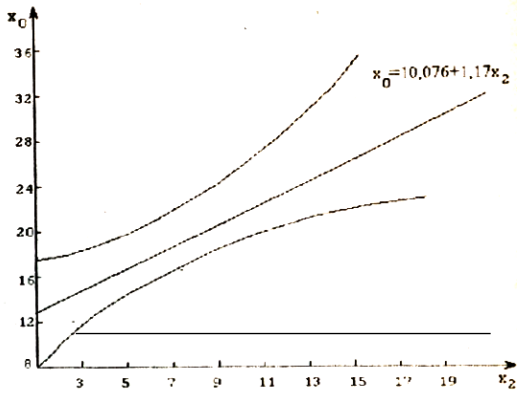


Рис. 6.1

Рассмотренная методика анализа качества оценивания среднеквадратической регрессии может быть (при сохранении условия нормальности распределения ошибок $\hat{\varepsilon}_j$) распространена и на многокомпонентные регрессии. Для этого достаточно, в частности, заменить в формулах (6.70) и (6.71) $t_{m-2,1-\beta/2}$ на $t_{m-k,1-\beta/2}$, где k – число оцениваемых коэффициентов.

6.2.7 Метод гребневой регрессии

Как отмечалось ранее, МНК-оценки коэффициентов регрессии, являются несмещенными. Однако, дисперсия этих оценок, вообще говоря, может быть значительной, что, естественно, вызывает и низкое качество оценивания x_0. В этих случаях можно ожидать, что качество выборочной регрессии повысится, если отказаться от рассмотрения только несмещенных оценок.

Но при прочих равных условиях дисперсия МНК-оценок коэффициентов регрессии тем больше, чем теснее корреляционные связи между предикторами. Так, для приведенной в примере 1 подразд. 6.2.6 двухкомпонентной регрессии

$$c_2^* = r_{0,2\,4}^* \frac{\delta_{0\,4}^*}{\delta_{2\,4}^*} = \frac{r_{0,2}^* - r_{0,4}^* r_{2\,4}^*}{1 - r_{2\,4}^{*2}} \frac{\delta_0^*}{\delta_2^*},$$

и, очевидно, при $r_{2\,4}^{*2} \approx 1$ даже небольшие случайные колебания оценок $r_{2,4}$ приводят к сильному разбросу МНК-оценок c_2. Поэтому задача отыскания оптимальных (в смысле минимума $M[\hat{\varepsilon}^2]$) коэффициентов регрессии (6.55) особенно важна при тесной корреляционной связи между привлекаемыми предикторами.

Для решения указанной задачи в последние годы предложен метод гребневой регрессии.

В основу метода гребневой регрессии положены следующие соображения. Будем характеризовать качество оценивания коэффициента c_i величиной $M[(\hat{c}_i^* - c_i)^2]$. Тогда использование смещенной оценки c_i^* вместо МНК-оценки

c_i^* (МНК) целесообразно, если

$M[(\hat{c}_i^* - c_i)^2] = D[\hat{c}_i^*] + \Delta^2 < D[\hat{c}_i^*(\text{мнк})]$,

где Δ - смещение оценки $\hat{c}_i^*$ относительно c_i.

Таким образом, при тесной коррелированности предикторов можно попытаться найти такие коэффициенты регрессии, которые при относительно небольшом смещении обладали бы значительно меньшей дисперсией, чем МНК-оценки.

Рассмотрим один из алгоритмов построения гребневой регрессии. Прежде всего для упрощения расчетов и анализа результатов введем стандартизованные переменные.

$y_{ij} = \frac{x_{ij} - \bar{x}_i}{\delta_{\tilde{x}_i}^*}$, i = 0, 1, …, n

и дополним таблицу исходных данных (табл. 6.1) n строками, содержащими нули и параметр k (табл. 6.5).

таблица 6.5

Номера реализаций	Стандартизованные значения y_i						
	y_0	y_1	y_2	…	y_i	…	y_n
1	y_{01}	y_{11}	y_{21}	…	y_{i1}	…	y_{n1}
2	y_{02}	y_{12}	y_{22}	…	y_{i2}	…	y_{n2}
…	…	…	…	…	…	…	…
j	y_{0j}	y_{1j}	y_{2j}	…	y_{ij}	…	y_{nj}
…	…	…	…	…	…	…	…
m	y_{0m}	y_{1m}	y_{2m}	…	y_{im}	…	y_{nm}
m+1	0	k	0	…	0	…	0
m+2	0	0	k	…	0	…	0
…	…	…	…	…	…	…	…
m+n	0	0	0	…	0	…	k

Оценка парных коэффициентов корреляции по дополненной таким образом выборке, очевидно, по модулю меньше, чем по исходной выборке, а оценки c_i смещены на величину Δ, определяемую параметром k.

Задаваясь значениями параметра k: δ, 2δ, … ($\delta > 0$), по данным табл. 6.5 найдем МНК-оценки коэффициентов регрессии.

$\tilde{y} = y_0 + d_1^* y_1 + d_2^* y_2 + \ldots + d_n^* y_n$,

для каждого значения параметра построим графики зависимости коэффициентов от k (гребневой след) и график зависимости $\tilde{\varepsilon}^2$ от k.

Оптимальным признается значение k, при котором оценка коэффициентов

регрессии «стабилизируется», а величина $\bar{\varepsilon}^2$ «не очень велика». Более конкретные рекомендации по выбору k можно дать, если значения $\bar{\varepsilon}^2$ определяются не только по основной («обучающей»), но и по дополнительной, не привлекавшейся при оценке коэффициентов «экзаменационной» выборке. В этом случае оптимальное значение k соответствует минимуму $\bar{\varepsilon}^2$ экз.

Пример 3. Воспользовавшись данными таблиц 6.2 и 6.3, построить линейную гребневую регрессию для прогноза максимальной температуры воздуха на ст. Омск x_0, по значениям x_1, x_2, x_4.

Целесообразность построения гребневой модели регрессии в рассматриваемом примере определяется наличием тесной корреляционной связи между предикторами $\hat{x}_1$ и $\hat{x}_2$ ($r_{\hat{x}_1\hat{x}_2}$ = 0,837).

Анализ начинаем с расчета для выборки 1 по данным табл. 6.2 и 6.3 стандартизированных значений x_0, x_1, x_2 и x_4 и заполнения верхней части табл. 6.6.

Для значений k, равных 1, 2, 3, …, найдем по данным табл. 6.6 МНК-оценки коэффициентов регрессии и рассчитаем средние квадратические ошибки регрессии по обучающей и экзаменационной выборкам.

Таблица 6.6

Номер реализаций	y_0	y_1	y_2	y_4
1	-	-	-	-
2	-	-	-	-
…	-	-	-	-
40	-	-	-	-
41	0	K	0	0
42	0	0	k	0
43	0	0	0	k

$$\tilde{y}_0 = d_0^* + d_1^* y_1 + d_2^* y_2 + d_4^* y_4$$

Представим результаты расчетов в виде графиков (рис. 6.2, 6.3). Анализ 6.2 показывает, что «поведение» оценок коэффициентов регрессии стабилизируется при $k > 3$. Но, как видно из рис. 6.3, при $k > 5$ происходит быстрый рост ошибок регрессии, определенных по обучающей выборке. Поэтому можно принять $k_{\text{опт}} \approx 4$. Этот вывод подтверждается и ходом кривой $\bar{\varepsilon}^2$, также показанной на рис. 6.3.

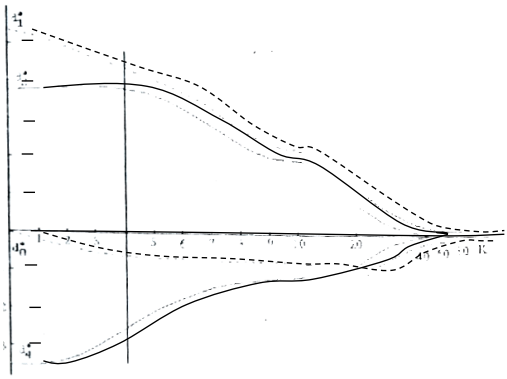


Рис. 6.2

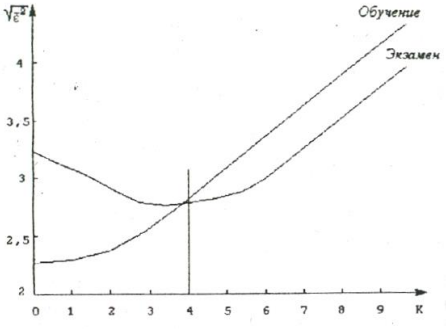


Рис. 6.3

Гребневая регрессия имеет вид:

$\tilde{y}_0$ = -0,047 + 0,384y_1 + 0,353y_2 – 0,235y_4

или, при переходе к исходным величинам,

$\tilde{x}_0$ = 11,910 + 0,469x_1 + 0,469x_2 – 0,436x_4.

Относительное уменьшение средней квадратической ошибки прогнозов по сравнению с аналогичной «обычной» регрессией равно по экзаменационной выборке 11,8%.

6.2.8 Робастные методы

При нормальном совместном распределении предиктанта $\hat{x}_0$ и предикторов $\hat{x}_1$, $\hat{x}_2$, …, $\hat{x}_n$ МНК-оценки являются подходящими значениями коэффициентов среднеквадратической регрессии.

Но, если условие нормальности совместного распределения $\hat{x}_0$ и $\hat{x}_{<n>}$ (а следовательно и $\hat{\varepsilon}$) не выполняется, МНК-оценки коэффициентов регрессии не

обеспечивают минимума $M[\hat{\varepsilon}^2]$ и можно попытаться найти оценки, более предпочтительные в указанном смысле.

Однако степень близости «истинного» распределения предиктанта и предикторов к нормальному, как правило, неизвестна. Поэтому желательно получить такие оценки, которые обеспечивали бы высокое качество регрессионных моделей при любых малых нарушениях предполагаемого закона.

Статистические модели, мало чувствительные к небольшим отклонениям от принятых при их разработке гипотез, называются робастными (устойчивыми).

Опыт прикладного регрессионного анализа показывает, что классические модели, основанные на МНК-оценивании коэффициентов регрессии, не отвечают требованию робастности. Дело в том, что при таком оценивании небольшие отклонения совместного распределения $\hat{x}_0$ и $\hat{x}_{<n>}$ от нормального приводят к появлению ошибок в «хвостах распределений» $\hat{\varepsilon}$.

На рис. 6.4 приводится гистограмма распределения ошибок выборочной регрессии

$$\tilde{x}_0 = 10{,}076 + 1{,}117x_2, \qquad (6.75)$$

рассмотренной в примере 2 подразд. 6.2.6. расчеты выполнены по данным табл. 6.2 (выборка 1).

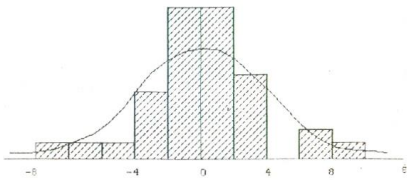


Рис. 6.4

Как видно на этом рисунке, при относительно общем неплохом согласии выборочного распределения с нормальным (при $M_{\hat{\varepsilon}} = \varepsilon = \overline{0}$ и $\delta_{\hat{\varepsilon}} = \delta_{\hat{\varepsilon}} = 3{,}519$) в «хвостах распределения» имеются существенные различия.

Но МНК-оценки очень чувствительны к ошибкам именно в «хвостах распределений» (так одна ошибка, равная 5, дает при МНК-оценивании такой же вклад в $\sum_j^- \varepsilon_j^2$, как 25 ошибок, равных 1), что в конечном итоге и обуславливает неустойчивость классических регрессионных моделей.

Сказанное дает основание ожидать, что для построения робастных регрессионных моделей следует уменьшить вклад «хвостов распределений» в суммарную ошибку регрессии. Ниже рассматривается один из алгоритмов построения робастной линейной регрессии, реализующий эту идею.

Оцениваются коэффициенты уравнения среднеквадратической регрессии

$$\hat{\tilde{x}}_0 = c_0 + c_1\hat{x}_1 + \ldots + c_n\hat{x}_n. \quad (6.76)$$

Как отмечалось в начале подразд. 6.2.5, МНК-оценки коэффициенты регрессии находятся из условия минимума суммы

$$\sum_{j=1}^{m}\left(\tilde{x}_{0j} - x_{0j}\right)^2 = \sum_j^m \varepsilon_j^2 = \sum_j^m |\varepsilon_j|\,|\varepsilon_j|, \quad (6.77)$$

то есть 2вклад» ошибки ε_j принимается равным ее абсолютной величине. При построении робастных регрессионных моделей для уменьшения вкладов крупных ошибок («хвостов распределений») Ньюбером предложено вместо (6.77) минимизировать сумму:

$$\sum_j^m(\rho_j), \quad (6.78)$$

$$\rho(\varepsilon) = \begin{cases} -h\left(\varepsilon + \dfrac{h}{2}\right), & \text{при } \varepsilon \le -h, \\ \dfrac{1}{2}\varepsilon^2 & \text{при } -h < \varepsilon < h, \\ h\left(\varepsilon - \dfrac{h}{2}\right) & \text{при } \varepsilon \ge h \end{cases}$$

График функции $\rho(\varepsilon)$ представлен на рис 6.5 и характеризуется линейной зависимостью ρ от ε вне интервала (-h, h).

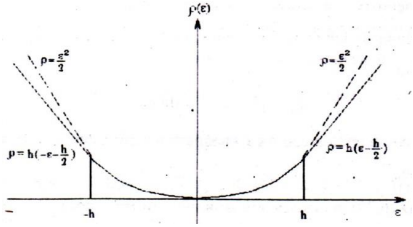


Рис. 6.5

Коэффициенты уравнения регрессии, обеспечивающие минимум суммы (6.78), являются решениями системы:

$$\left.\begin{aligned} h\left(\sum_{\varepsilon\ge h} x_{ij} - \sum_{\varepsilon\le h} x_{ij}\right) + \sum_{(-h<\varepsilon<h)} \varepsilon_j x_{ij} = 0 \;\; i = \overline{1,n} \\ h(n_+ - n_-) + \sum_{(-h<\varepsilon<h)} \varepsilon_j = 0, \end{aligned}\right\} \quad (6.79)$$

где $\varepsilon_j = c_0^* + c_1^* x_{ij} + \ldots + c_n^* x_{nj} - x_{0j}$,

n_+ и n_- – числа ошибок $\varepsilon_j \ge h$ и $\varepsilon_j \le -h$ соответственно.

Параметр робастности h задается в зависимости от распределения ошибок выборочной регрессии. Очевидно, при $h > |\varepsilon|_{max}$ будут получены те же оценки, что и методом наименьших квадратов, а при $h > 0$ будет минимизирована сумма

абсолютных ошибок.

Судя по распределению ошибок регрессии (6.80) (рис. 6.4) существенные отличия этого распределения от нормального наблюдается при $|\varepsilon| > 1…2$. Поэтому можно ожидать, что для построения робастной регрессии параметру h следует придать значения, близкие к единице. Этот вывод подтверждается анализом зависимости от h средней квадратической ошибки прогнозов, составленных для экзаменационной выборки (рис. 6.6).

Наилучшие результаты получены для h = 0,60. Соответствующее уравнение регрессии

$$\tilde{x}_0 = 10{,}005 + 1{,}089x_2, \qquad (6.80)$$

обеспечивает уменьшение средней квадратической ошибки таких прогнозов примерно на 3%.

Система (6.79) решается интерационными методами. В качестве исходных значений коэффициентов регрессии используются МНК-оценки. (Для построения регрессии (6.80) потребовалось 27 интераций).

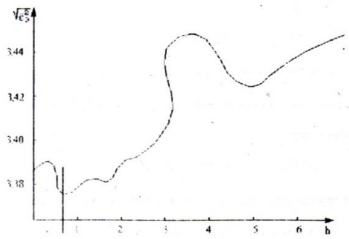


Рис. 6.6

Ниже рассматривается пример построения многокомпонентный робастной регрессии.

Пример 4. Пользуясь данными табл. 6.2, 6.3 и 6.4, построить робастную регерссию для прогноза максимальной температуры воздуха (x_0) по значениям предикторов x_1, x_2 и x_4.

Требуется оценить коэффициенты уравнения регрессии

$\hat{\tilde{x}}_0 = c_0 + c_1\hat{x}_1 + c_2\hat{x}_2 + c_4\hat{x}_4$.

Начнем с нахождения МНК-оценок. Подставив в уравнения (6.61) соответствующие оценки парных коэффициентов корреляции из табл. (6.4) и средних квадратических отклонений из табл. (6.3), получим системуб

$$\left.\begin{aligned} 6{,}558\ \ 0838 &= 5{,}366c_1 + 4{,}932\ 0{,}837c_2 - 3{,}537\ 0{,}026c_4, \\ 6{,}558\ \ 0{,}840 &= 5{,}366\ 0{,}837c_1 + 4{,}932c_2 - 3{,}537\ 0{,}106c_4, \\ 6{,}558\ \ 0{,}384 &= -5{,}366\ \ 0{,}026c_2 - 4{,}932\ \ 0{,}106c_2 + 3{,}537c_4. \end{aligned}\right\}$$

Решение системы c_1= 0,639; c_2= 0,487; c_4 = -0,615. Теперь с учетом (6.58) и значений $M^*_{\hat{x}_i}$ (табл. 6.3) найдем

$c_0 = 19{,}375 - (0{,}639\ \ 5{,}366 + 0{,}487\ \ 4{,}932 - 0{,}615\ \ 3{,}537) = 10{,}964.$

Таким образом, выборочная регрессия при МНК-оценивании ее коэффициентов принимает вид:

$$\tilde{x}_0 = 10{,}964 + 0{,}639x_1 + 0{,}487x_2 - 0{,}615x_4 \qquad (6.81)$$

Для определения целесообразности перехода к робастному уравнению рассчитаем по полученной формуле значения $\tilde{x}_0$ для всех вариантов выборки 1 (табл.6.2) и построим гистограмму распределения ошибок, по которой отметим значения плотности, соответствующие нормальному закону (рис. 6.7).

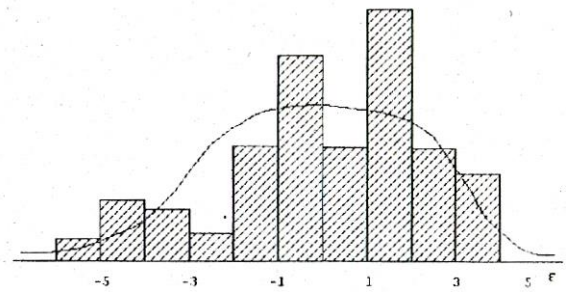


Рис.6.7

Анализ рис. 6.7 показывает, что отличия выборочного распределения ошибки $\hat{\varepsilon}$ от нормального имеет место не только в «хвостах распределения», но и в его центральной части. Поэтому выбор оптимального значения h по обучающей выборке затруднителен. На рис. 6.8 показана зависимость от h средней квадратической ошибки прогнозов, определенной для экзаменационной выборки (выборка 2 в табл. 6.2).

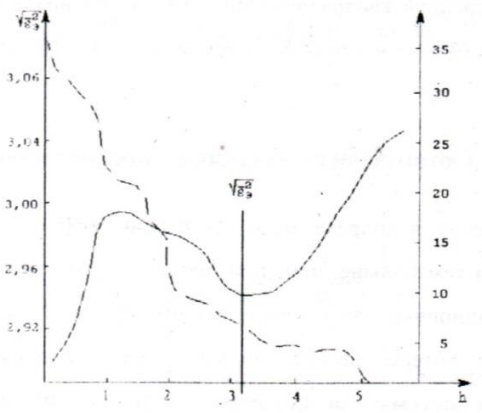


Рис. 6.8

Частный минимум $\bar{\varepsilon}_8^2$ был получен при h ≈ 3, что в общем согласуется с распределением $\hat{\varepsilon}$ на рис. 6.7. общее число ошибок (в обучающей выборке), превосходящих по абсолютной величине h$(n_+ + n_-)$ оказалось равным 13 – около 30% от объема выборки. Однако глобальный минимум $\bar{\varepsilon}_8^2$ был достигнут при h > $|\varepsilon|_{min}$, то есть, фактически, при минимизации по обучающей выборке не $\sum_j^{40} \varepsilon_j^2$, а $\sum_j^{40} |\varepsilon_j|$. Иначе говоря, оказалось, что при рассматриваемом распределении ошибок для уменьшения M$[\hat{\varepsilon}^2]$ коэффициенты уравнения регрессии следует оценивать не методом наименьших квадратов, а методом наименьших модулей.

Уравнение робастной регрессии при h = 0.1$(n_+ + n_-)$ = 36 имеет вид:

$$\tilde{x}_0 = 10{,}552 + 0{,}656x_1 + 0{,}486x_2 - 0{,}530x_4 \qquad (6.82)$$

Снижение средней квадратической ошибки прогнозов при использовании уравнения (6.82) вместо (6.81) составило по экзаменационной выборке 4,6%.

6.2.9 Ортогонализация систем предикторов

Как отмечалось в подразд. 6.2.7, дисперсии МНК-оценок коэффициентов регрессии тем больше, чем, при прочих равных условиях, теснее взаимные корреляционные связи между предикторами. С этой точки зрения представляют интерес регрессионные модели, предусматривающие ортогонализацию системы предикторов. В метеорологической практике такая ортогонализация осуществляется, как правило, методами компонентного анализа, подробно рассмотренными в подразд. 6.4.2.

При компонентном анализе обычно сначала вводятся нормализованные случайные величины:

$$\hat{y}_i = \frac{\hat{x}_i - M_{\hat{x}_i}}{\delta_{\hat{x}_i}},\ i = 1, 2, \ldots, n \qquad (6.83)$$

затем вектор-предиктор $\hat{Y}_{<n>} = \langle \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n \rangle$ представляется в виде $\hat{Y}_{<n>} = \langle \hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n \rangle$, где $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n$ - новые составляющие вектора $Y_{<n>}$, не коррелирующие друг с другом и линейно связанные с исходными предикторами (главные компоненты):

$$\left.\begin{aligned} \hat{z}_1 &= a_1^{(1)}\hat{y}_1 + a_2^{(1)}\hat{y}_2 + \ \ldots + a_n^{(1)}\hat{y}_n, \\ \hat{z}_2 &= a_1^{(2)}\hat{y}_1 + a_2^{(2)}\hat{y}_2 + \ \ldots + a_n^{(2)}\hat{y}_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \hat{z}_n &= a_1^{(n)}\hat{y}_1 + a_2^{(n)}\hat{y}_2 + \ \ldots + a_n^{(n)}\hat{y}_n, \end{aligned}\right\} \qquad (6.84)$$

Коэффициенты $a_i^{(k)}$ в (6.84) находятся из условия6

$$\left.\begin{aligned} A_{<n>}^{(1)} &= \langle a_1^{(1)}, a_2^{(1)}, \ldots, a_n^{(1)} \rangle, \\ A_{<n>}^{(2)} &= \langle a_1^{(2)}, a_2^{(2)}, \ldots, a_n^{(2)} \rangle, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ A_{<n>}^{(n)} &= \langle a_1^{(n)}, a_2^{(n)}, \ldots, a_n^{(n)} \rangle, \end{aligned}\right\} \qquad (6.85)$$

где $A^{(1)}_{<n>}$, $A^{(2)}_{<n>}$, …, $A^{(n)}_{<n>}$ – собственные векторы корреляционной матрицы для системы исходных предикторов.

Уравнение среднеквадратической регрессии

$$\hat{\tilde{y}}_0 = d_1\hat{z}_1 + d_2\hat{z}_2 + \cdots + d_n\hat{z}_n \quad (6.86)$$

с учетом (6.63) и (6.84) легко приводится к окончательному виду:

$$\hat{\tilde{x}}_0 = c_1\hat{x}_1 + c_2\hat{x}_2 + c_n\hat{x}_n. \quad (6.87)$$

При построении выборочных моделей параметра $M_{\hat{x}_i}$, $\delta_{\hat{x}_i}$, $r_{\hat{x}_i,\hat{x}_1}$ заменяются их статистическими оценками.

Пример 5. Основываясь на данных табл. 6.2, 6.3 и 6.4 и выполнив ортогонализацию системы предикторов методом главных компонент, построить линейную среднеквадратическую регрессию для прогноза максимальной температуры воздуха $\hat{x}_0$.

Найдем собственные числа и собственные векторы корреляционной матрицы (табл. 6.4). в табл. 6.7 собственные векторы пронумерованы в порядке убывания собственных чисел. Полученные результаты показывают, что учет первых четырех главных компонент обеспечивают снятие более выборочной дисперсии $\hat{X}_{<8>}$:

1/8(2,979 + 1,857 + 1,040 + 0,953) 100 = 85,3

Таблица 6.7

N	Собственные числа	Собственные векторы							
1	2,979	<-0,238	-0,286	-0,453	0,459	0,111	0,459	0,345	0,322>
2	1,857	<-0,605	-0,589	0,214	-0,277	-0,183	-0,218	-0,141	0,250>
3	1,040	<0,110	-0,047	-0,366	0,355	0,152	-0,369	-0,709	0,252>
4	0,953	<0,064	0,100	-0,206	0,102	-0,958	0,088	-0,091	0,005>
5	0,673	<0,322	0,186	0,255	-0,190	-0,030	0,065	0,079	0,866>
6	0,250	<-0,183	0,268	-0,366	-0,052	-0,023	-0,694	0,513	0,111>
7	0,151	<-0,526	0,518	0,445	0,494	-0,14	0,015	-0,086	0,068>
8	0,097	<-0,383	0,429	-0,419	-0542	0,108	0,333	-0,272	0,058>

Обратившись к табл. 6.2, 6.3 и 6.7 и воспользовавшись формулами (6.83) и (6.84), найдем первое значение первой главной компоненты:

$$z_{11} = -\frac{10-10{,}225}{5{,}336}0{,}238 - \frac{8-8{,}325}{4{,}932}0{,}286 - \frac{9-4{,}975}{3{,}051}0{,}453 + \\ + \frac{0-3{,}550}{3{,}537}0{,}459 + \frac{9-5{,}425}{2{,}678}0{,}111 + \frac{0-4{,}000}{3{,}693}0{,}459 + \\ + \frac{0-5{,}350}{3{,}676}0{,}345 + \frac{60-80{,}650}{10{,}076}0{,}322 = -2{,}539.$$

Аналогично определим первое значение второй главной компоненты z = 0,309, вторые значения этих компонент: z_{12} = -1,722, z_{22} = 0,251 и т. д.

МНК-оценки коэффициентов регрессии (6.86) найдем по обучающей выборке (выборка 1 в табл. 6.2), а прогностическую эффективность моделей будем характеризовать средними квадратическими ошибками прогнозов максимальной температуры по экзаменационной выборке (выборка 2).

Для отбора предикторов воспользуемся процедурой их последовательного включения (см. г. 9).

На рис. 6.9 приведены значения средних квадратических ошибок регрессии (6.87) для обучающей и экзаменационной выборок при различных состояниях главных компонент.

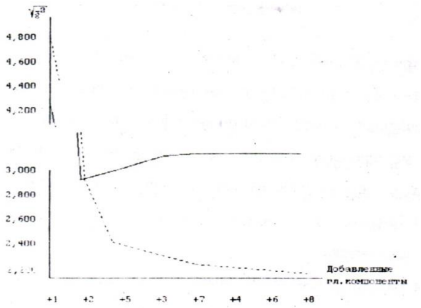


Рис. 6.9

Лучшая из рассмотренных моделей включает две главные компоненты z_1 и z_2. Соответствующее уравнение регрессии имеет вид:

$y_0 = -0{,}391z_1 - 0{,}432z_2,$

для исходных переменных

$$x_0 = 23{,}491 + 0{,}433x_1 + 0{,}487x_2 + 0{,}182x_3 - 0{,}111x_4 + \\ + 0{,}087x_5 - 0{,}151x_6 - 0{,}132x_7 - 0{,}152x_8. \qquad (6.88)$$

Средняя квадратчиеская ошибка регрессии (6.88) для экзаменационной выборки составила 2,830, в то время как для лучшей регрессии, построенной непосредственно по исходным предикторам, она равна 3,064. Таким образом, ортогонализация системы предикторов позволила в рассмотренном примере снизить ошибку прогнозов примерно на 7%.

6.3 Дискриминантный и кластерный анализ

6.3.1 Общие сведения о дискриминантном и кластерном анализе

Дискриминантный анализ является одним из видов статистического анализа связей и применяется в тех ситуациях, когда результат (предиктант) является качественной характеристикой, а факторы (предикторы) – количественные. В метеорологии дискриминантный анализ выделился в самостоятельный раздел статистических методов обработки метеорологической информации в 70-е годы,

хотя в неявном (описательном) виде с этими методами работают с начала века.

Сам термин «дискриминантный» анализ, используемый в метеорологической практике, не совсем точно отражает смысл проблемы, которая впервые была поставлена кибернетиками как содержательная задача в конце 50-х годов. Дискриминантный анализ с точки зрения кибернетиков является составной частью теории распознания образов.

Теория распознания образов возникла в связи с необходимостью обучения машин распознаванию объектов, т. е. разработкой алгоритмов позволяющих однозначно идентифицировать различные объекты. Сложность этой задачи можно пояснить на простом примере. Представим себе, что у нас есть стол и есть табурет. Требуется разработать такой алгоритм принятия решения, по которому ЭВМ однозначно определила бы, что стол является столом, а табурет – табуретом. Для человека эта задача никакой сложности не представляет, так как он представляет, что стол – это то, за чем сидят, а табурет – на чем сидят. Причем, если задать дополнительную информацию – стол обеденный, стол письменный, то в нашем мозгу возникает образ такого стола, ассоциирующийся с тем, что мы видели раньше. Аналогично обстоит дело и с табуретом.

К сожалению, у ЭВМ ассоциативное мышление отсутствует и она не может себе «представить» ни стола, ни табурета. Следовательно, ей нужно задать такие количественные характеристики и такие правила разделения значений этих характеристик, по которым она могла бы однозначно отнести эти объекты к одному из двух возможных классов «стол» и «табурет». Вот здесь и возникает основная проблема. Такими характеристиками могут служить наличие крышки и количество ножек, размеры крышки и размеры ножек. К сожалению, но этим характеристикам провести идентификацию объекта не удается, так как у стула, и у табурета по одной крышке и по определенному количеству ножек. Есть столы и табуреты с четырьмя ножками, есть с тремя, есть с одной. С размерами также дело обстоит неважно, так, столы и табуреты есть большие и маленькие, есть «взрослые» и «детские» и др.

Таким образом, задача обучения машин распознаванию образов – теория распознавания образов – проста только на первый взгляд. Она также как и другие разделы статистических методов обработки информации имеет свои специальные особенности и свои методы решения.

В этой теории можно выделить две основные задачи принципиально отличающиеся друг от друга.

К первому типу отнесем задачу классификацию, суть которой состоит в том, что в некотором известном пространстве признаков требуется построить поверхность, с одной стороны от которой были объекты одного класса, а с другой

стороны – второго класса (если классов будет больше двух, то и количество поверхностей соответственно увеличится). Такие задачи в теории распознавания образов получили название задач кластеризации, а вся совокупность таких задач – кластерного анализа. Примером кластерного анализа в синоптической практике является анализ воздушных масс и атмосферных фронтов. Задача анализа воздушных масс состоит в том, чтобы определить по ряду измеряемых на метеорологических станциях величин наличие неоднородных типов «погод» и количество таких неоднородностей. Задача анализа атмосферных фронтов состоит в определении положения поверхностей, разделяющих эти неоднородности.

Ко второму типу задач теории распознавания образов отнесем задачи дискриминации (разделения). Эти задачи отличаются от задач кластеризации тем, что классы уже известны и требуется новую «неизвестную» точку отнести к одному из этих неизвестных классов. Совокупность методов решения таких задач называют дискриминантным анализом. Примером использования методов дискриминантного анализа в синоптической практике может служить задача прогноза опасных явлений погоды – по известному измеренному комплексу метеорологических величин требуется определить возможность возникновения грозы или ее отсутствия.

Таким образом, различие между дискриминантным и кластерным анализом состоит в том, что в дискриминантном анализе характеристики классов и их количество известны, а в кластерном их требуется найти. Сначала выполняется этап кластеризации, а затем – дискриминации.

В метеорологии сложилось так, что под дискриминантным анализом понимают совокупность задач кластеризации и дискриминации т. е. теорию распознавания образов в целом называют дискриминантным анализом. Мы в дальнейшем также будем придерживаться этой терминологии, за исключением особо оговоренных случаев.

Рассмотрим суть дискриминантного анализа на простом примере. Пусть известно, что при некоторых сочетаниях температуры и влажности туман возникает, а при некоторых – нет. Требуется найти ту область сочетаний температуры и влажности, при которой возникает туман. Для решения этой задачи выберем систему координат таким образом, чтобы по одной оси откладывались значения влажности R, а по другой – температуры T (рис. 6.10).

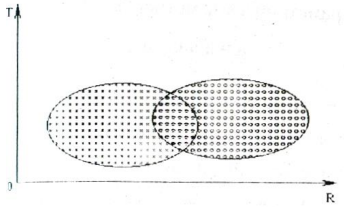


Рис. 6.10

Рассмотрим результаты наблюдений за этим явлением по архивной выборке. Каждому сочетанию температуры и влажности в нашей системе координат будем ставить в соответствие точку, причем если сочетанию соответствовало возникновение тумана через некоторый промежуток времени, то ее обозначим крестиком, а если нет – ноликом. В результате на поле нашего графика образуется две области: область крестиков и область ноликов. Если очертить эти области замкнутыми кривыми, то легко видеть, что эти области пересекаются (в случае отсутствия пересечения проблем с дискриминацией нет), т. е. при некоторых сочетаниях температуры и влажности туман может возникнуть, а может и не возникнуть. Требуется разделить некоторой линией крестики и нолики таким образом, чтобы количество ошибочно классифицированных крестиков и ноликов было минимальным.

Если пространство факторов больше двух, то суть дела не изменяется, просто вместо линии классы будут разделяться поверхностью или гиперповерхностью.

В случае появления новой точки, требуется определить, с какой стороны от полученной линии классы будет находиться эта точка.

В зависимости от того, известны параметры классов или нет, все методы дискриминантного анализа делятся на две группы: параметрические и непараметрические.

6.3.2 Непараметрические методы дискриминантного анализа

Непарметрические методы дискриминантного анализа используются в ситуациях, когда неизвестны законы и, соответственно параметры законов распределения классов (рис. 6.10). в таких ситуациях требуется построить правило отнесения новой «неизвестной» точки к одному из имеющихся классов (рис. 6.11). пусть, например, в результате испытаний получены два сочетания факторов x_1 и x_2, причем одному сочетанию соответствует «крестик» и второму «нолик». Требуется выяснить, к какому классу относится третья «неизвестная» точка. Из логических соображений эту точку следует отнести к тому классу, к которому она «ближе», т. е. если эта точка ближе к крестику, то ее следует отнести к крестикам,

в противном случае – к ноликам.

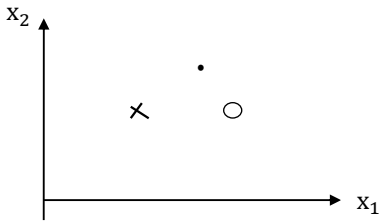


Рис.6.11

Таким образом, если мы сможем рассчитать расстояние между точками в факторном пространстве, то задачу можно решить с определенной степенью точности. В связи с тем, что факторы имеют различную размерность, не связанную, как правило, с линейной размерностью, в качестве меры близости используют евклидово расстояние, которое обычно обозначают через r. Смысл евклидова расстояния поясним с помощью рис. 6.12.

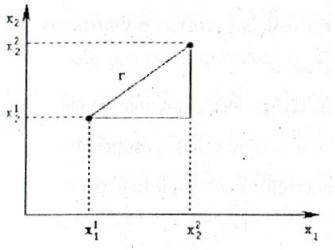


Рис.6.12

В пространстве факторов x_1, x_2 заданы две точки с координатами (x_1^1, x_2^1), и (x_1^2, x_2^2). Требуется найти расстояние между ними. Очевидно, что если известны координаты точек, то можно рассчитать катеты соответствующего треугольника и по теореме Пифагора евклидово расстояние будет определяться выражением

$r = \sqrt{(x_2^2 - x_2^1)^2 + (x_1^2 - x_1^1)^2}$.

Если количество факторов увеличиться до трех, то под корнем будет три слагаемых, до четырех – четыре и т. д. Поэтому в общем виде евклидово расстояние между двумя точками в n-мерном пространстве факторов будет записано в виде

$$r(x_1,x_2) = \sqrt{\sum_{i=1}^{n}(x_i^2 - x_i^1)^2}, \qquad (6.89)$$

где x_i^1 – i-ая координата первой точки,

x_i^2 – i-ая координата первой точки.

На рис.6.12 представлено всего два сочетания факторов x_1 и x_2. При статистической обработке таких сочетаний будет неизмеримо больше и сразу же возникает вопрос: до чего считать расстояние от новой точки.

Все методы непараметрического дискриминантного анализа и отличаются друг от друга объектом, до которого считается расстояние. Рассмотрим некоторые из этих методов.

1. Метод эталонов. Пусть известны две области: крестиков и ноликов в пространстве факторов x_1 и x_2 (рис. 6.13), которые обозначим соответственно w_1 и w_2. Каждая из этих областей имеет свой центр тяжести 0_1 и 0_2. Назовем эти центры эталонами.

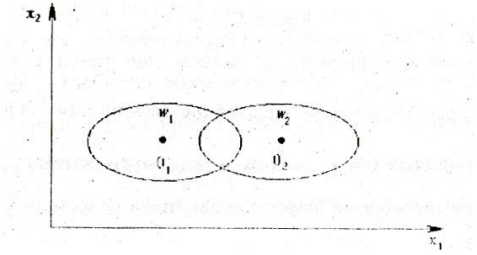


Рис. 6.13

Суть метода эталонов состоит в том, что при появлении новой «неизвестной» точки рассчитывается расстояние до этих центров. Точку следует отнести к классу w_1, если $r(x, 0_1) < r(x, 0_2)$. В противном случае точка относится к классу w_2. Если вспомнить, что в теории вероятностей аналогом центра тяжести является математическое ожидание, то для реализации метода эталонов необходимо рассчитать математическое ожидание точек, принадлежащих классу w_1 и математическое ожидание точек, принадлежащих классу w_2 по формуле (6.89).

Таким образом, задача построения метода эталонов сводится к задаче нахождения центров классов и дальнейшему расчету расстояния от новых точек до этих центров.

Метод эталонов является очень простым, так как математические ожидания классов можно рассчитать один раз и в дальнейшем расчеты вести только для двух точек 0_1 и 0_2. Однако он обладает очень существенным недостатком. На рис. 6.14 показан случай, когда в одномерном пространстве признаков разделение крестиков и ноликов происходит идеально, а в двумерном – один из ноликов по методу эталонов должен быть отнесен к классу крестиков, т. е. возникает ошибка дискриминации. Из этого примера видно, что увеличение пространства

признаков в случае метода эталонов не приводит к улучшению качества разделения, хотя для других методов это является характерным признаком дискриминации.

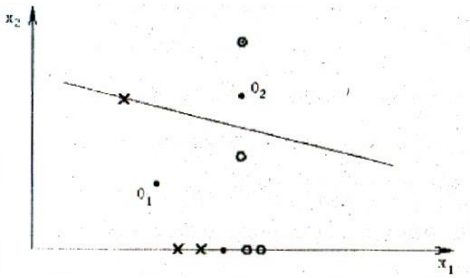


Рис. 6.14

2. Метод ближайшей точки (ближайшего соседа). Суть этого метода также очень проста. Нужно измерить расстояние от новой («неизвестной») точки до всех точек соответствующих классов и среди этих рассчитанных расстоянии выбирать минимальное. Неизвестную точку нужно отнести к тому классу, к которому относится ближайшая точка.

Несмотря на внешнюю простоту, метод ближайшего соседа оказывается часто эффективнее метода эталонов по следующим причинам. Если количество точек устремить к бесконечности, то вероятность того, что ближайшая точка относится к своему классу, будет стремиться к истинной вероятности, а так как это точная вероятность того, что природа находится в указанном состоянии, то правило ближайшего соседа эффективно согласует вероятности с реальностью.

3. Метод k ближайших соседей. Этот метод является явным расширением правила ближайшего соседа. Как видно из названия, это правило классифицирует неизвестную точку, присваивая ему метку (крестик или нолик), наиболее часто представляемую среди k ближайших точек.

Реализация алгоритма k ближайших соседей в следующем. Из неизвестной точки с радиусом R проводим окружность и подсчитаем сколько крестиков и ноликов попало в окружность. Если больше оказалось крестиков, то неизвестную точку необходимо отнести к крестикам, в противном случае – к ноликам.

4. Линейные разделяющие поверхности. Пусть известны два класса w_1 и w_2 в пространстве x_1, x_2 и известны математические ожидания (центры) классов. Требуется разделить эти классы прямой линией таким образом, чтобы было достигнуто минимальное число случаев попадания точек в «чужие» классы.

Эта задача часто решается с помощью метода эталонов. Для ее решения центры классов соединяют прямой линией и через середину этой линии восстанавливают перпендикуляр (рис. 6.15). естественно, что в этом случае все точки,

лежащие слева от перпендикуляра, будут ближе к центру 0_1, а справа – к центру 0_2. Поэтому левая полуплоскость должна быть отнесена к классу w_1, а правая к классу w_2. Такое разделение будет оправдано, если классы w_1 и w_2 имеют одинаковую компактность, характеризуемую дисперсиями D_{w_1} и D_{w_2} или среднеквадратическими отклонениями δ_{w_1} и δ_{w_2}. Таким образом, если $\delta_{w_1} = \delta_{w_2}$, то данный алгоритм позволяет решить поставленную задачу.

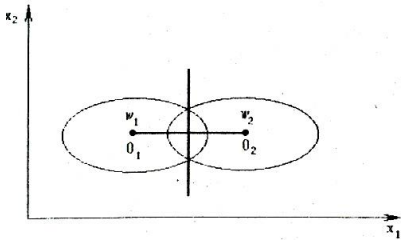


Рис. 6.15

Уравнение этого перпендикуляра запишется как уравнение прямой в пространстве x_1, x_1 с помощью выражения

$$K = a_0 + a_1 x_1 + a_2 x_2. \quad (6.90)$$

Если пространство признаков больше двух, то вместо разделяющей (дискриминантной) линии мы получим дискриминантную поверхность или гиперповерхность в виде

$$K = a_0 + a_1 x_1 + \cdots + a_n x_n, \quad (6.91)$$

где $x_1, x_2, \ldots, x_n$ – факторы;

$a_0, a_1, \ldots, a_n$ – коэффициенты дискриминантной поверхности.

В случае, когда $\delta_1 \neq \delta_2$ (компактность классов разная), как это представлено на рис. 6.16, построенный указанным выше способом перпендикуляр не будет характеризовать оптимальное положение разделяющей поверхности, т. к. большое количество точек, принадлежащих классу w_2, окажется ошибочно отнесенным к классу w_1. Из рис. 6.16 видно, что эта поверхность должна проходить левее построенного перпендикуляра – через середину области пересечения классов, т. е. уравнение разделяющей поверхности будет отличаться от уравнения перпендикуляра своими коэффициентами $a_0, a_1, \ldots, a_n$.

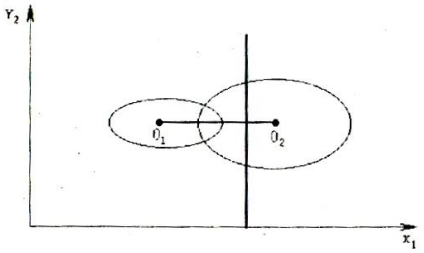


Рис. 6.16

Возникает задача нахождения коэффициентов уравнения дискриминантной поверхности, обеспечивающей наилучшее разделение классов. Так как понятие «наилучшее разделение» связано, прежде всего, с видом критерия разделимости, то и положение дискриминантной поверхности будет изменяться при переходе от одного критерия к другому. Таким образом, задача нахождения оптимального положения дискриминантной поверхности сводится к задаче нахождения экстремума соответствующего критерия. Такие задачи решаются с помощью методов градиентного спуска, суть которых можно объяснить с помощью метода «слепого альпиниста».

Если альпинисту завязать глаза и дать задание подняться на вершину, то его действия будут сведены к ощупыванию поверхности вокруг себя. Очевидно, что он должен сделать шаг в ту сторону, где наблюдается максимальная крутизна подъема, т. е. выбрать направление движения по максимальному градиенту. В новой точке он опять должен ощупать вокруг себя и следующий шаг совершить в сторону максимального градиента и т. д. Таким образом, проверяя на каждом шаге крутизну подъема, альпинист выйдет на вершину горы.

Данный способ позволит найти экстремумы функции в случае, если вершина одна. Если же их несколько, то таким путем можно попасть на вторичный максимум, а не на главный.

Для нахождения главного максимума задачу решают с помощью изменения величины шага. На первом этапе шаг выбирают большой (исходя из условий конкретной задачи) и ищут район, в котором находится абсолютный максимум. На втором этапе поиск экстремума осуществляется в пределах полученного района с меньшим шагом и т. д. до тех пор, пока с нужной точностью не будет найден экстремум функции. В настоящее время разработано достаточно много алгоритмов, позволяющих наилучшим образом реализовать спуск, наискорейший спуск, изменение масштабов. Давидона-Флетчера-Пауэлла и др. эти методы достаточно подробно описаны в соответствующей литературе и здесь рассматриваться не будет.

Используя указанные методы, можно найти коэффициенты уравнения (6.91), обеспечивающие такое положение разделяющей поверхности, при котором достигается экстремум соответствующего критерия.

Пусть, например, имеются две пересекающиеся области крестиков и ноликов. Требуется эти области разделить прямой линией таким образом, чтобы было достигнуто минимальное количество ошибочных попаданий точек в «чужой» класс. На первом шаге примем, что $a_1 = a_2 = \ldots = a_n = 1$, а коэффициент a_0 начнем изменять с некоторым шагом до тех пор, пока не найдем минимум ошибок разделения. Значение a_0, при котором будет обеспечен минимум, зафиксируем и начнем изменять значения коэффициента a_1 и т. д. После нахождения значений всех коэффициентов процедура повторяется и так до тех пор пока не будет достигнут нужный результат.

5. Метод потенциальных функций. Толчком к созданию метода потенциальных функций послужило следующее обстоятельство: если векторы наблюдений x_1 представить как точки некоторого пространства и в эти точки поместить заряды $+q_1$, если x_i помечено символом w_1 и $-q_i$, если x_i помечено символом w_2, то возможно, функцию, описывающую распределение электростатического потенциала в таком поле, можно будет использовать в качестве разделяющей функции. Если потенциал точки x, создаваемый единичным зарядом, находящимся в точке x_i, равен K(x, x_i), то потенциал, создаваемый n зарядами в точке x, определяется выражением

$$g(x) = \sum_{i=1}^{n} q_i K(x, x_i). \qquad (6.92)$$

Потенциальная функция K(x, x_i), используемая в классической физике, обратно пропорциональна расстоянию между точками. Имеется и много других функций, которые могут быть использованы для нашей цели. Наиболее часто используется такая потенциальная функция, которая имеет максимум при x = x_i и монотонно убывает до нуля при r→ ∞.

Пусть имеется множество из n случаев, а разделяющая функция сформирована в соответствии с выражением (6.92). предположим далее, что при проверке обнаружено, что некоторый случай, например x_k, посредством функции q(x) классифицируется с ошибкой. Попробуем исправить ошибку, изменив немного величину заряда q_k. Предположим, что значение q_k увеличивается на величину единичного заряда, если x_k помечено символом w_1, и уменьшается на такую же величину, если x_k помечено символом w_2. Если обозначить значение разделяющей функции после коррекции через g′(x), то алгоритм формирования данной функции может быть записан в следующем виде:

$$g'(x) = \begin{cases} g(x) + K(x, x_k), & \text{если } x_k \text{ обозначен } w_1 \text{ и } g(x_k) \le 0 \\ g(x) - K(x, x_k), & \text{если } x_k \text{ обозначен } w_2 \text{ и } g(x_k) \ge 0 \\ g(x) & \text{в противном случае} \end{cases}$$

Проводя такую корреляцию некоторое число раз можно добиться идеального разделения точек. Однако, следует заметить, что идеальное разделение точек на исходной выборке, как правило, при переходе к контрольным выборкам не выполняется. На контрольных выборках результаты получаются намного хуже. Поэтому стремятся достичь такого качества разделения, которое обеспечивалось бы и при переходе к другим выборкам.

6.3.3 Параметрические методы дискриминантного анализа

Параметрические методы дискриминантного анализа используются в тех случаях, когда известны параметры законов распределения соответствующих классов.

Рассмотрим суть этих методов на примере. Из курса физики атмосферы известно, что грозы возникают в тех случаях, когда в атмосфере в достаточно большом слое имеется положительная энергия неустойчивости. Определяют эту энергию неустойчивости с помощью аэрологической диаграммы. Проведем статистическую обработку всех случаев, когда наблюдалась гроза, и случаев, когда она не наблюдалась. В результате получим два закона распределения, представленных на рис. 6.17. Для первого закона каждому значению энергии неустойчивости x поставили в соответствие вероятность отсутствия грозы, а для второго – вероятность наличия грозы. Обозначим класс со случаями «без грозы» через w_1, а «с грозой» - w_2. Пусть $p(x/w_j)$ – условная плотность распределения величины x в состоянии w_j, т. е. функция плотности распределения случайной величины $\hat{x}$ при условии, что состояние природы есть w_j. В этом случае различие между $p(x/w_1)$ и $p(x/w_2)$ отражает факт наличия или отсутствия грозы.

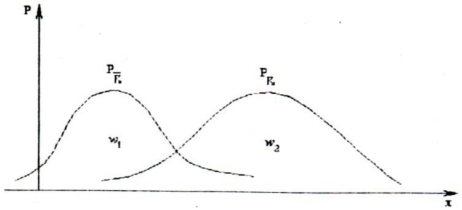


Рис.6.17

Допустим, что известны как априорные вероятности $p(w_j)$, так и условные

плотности $p(x/w_j)$. Предположим далее, что мы вычисляем значение энергии неустойчивости и находим, что оно равно x. ответ на вопрос, в какой мере это вычисление повлияет на наше представление об истинном состоянии природы, дает правило Байеса:

$$p(w_j/x) = \frac{p(x/w_j)p(w_j)}{p(x)}, \tag{6.93}$$

$$p(x) = \sum_{j=1}^{2} p(x/w_j)\, p(w_j). \tag{6.94}$$

Правило Байеса показывает, как наличие вычисленной величины $\hat{x}$ позволяет из априорной вероятности $p(w_j)$ получить апостериорную вероятность $p(w_j/x)$. Если вычислении получено значение $\hat{x}$, для которого $p(w_1/x)$ больше, чем $p(w_2/x)$, то естественно склониться к решению, что истинное состояние природы есть w_1. Аналогично, если $p(w_2/x) > p(w_1/x)$, то естественно склониться к выбору w_2. Чтобы обосновать этот выбор, вычислим вероятность ошибки при принятии решения, когда наблюдалось определенное значение $\hat{x}$:

$$p(\text{ошибка}/x) = \begin{cases} p(w_1/x), \text{если принято решение } w_2 \\ p(w_2/x), \text{ если принято решение } w_1 \end{cases}$$

Очевидно, что в каждом из случаев, когда одно и то же значение x, вероятность ошибки можно свести к минимуму, принимая решение w_1, если $p(w_1/x) > p(w_2/x)$, и w_2, если $p(w_2/x) > p(w_1/x)$. разумеется, мы не можем дважды наблюдать точно то же самое значение величины $\hat{x}$ и поэтому возникает вопрос: будет ли это правило минимизировать среднюю вероятность ошибки? Можно ответить однозначно – да, так как средняя вероятность ошибки определяется выражением

$$p(\text{ошибка}) = \int_{-\infty}^{\infty} p(\text{ошибка}/x)p(x)dx,$$

и если для каждого $\hat{x}$ вероятность $p(\text{ошибка}/x)$ достигает наименьшего значения, то и интеграл должен быть минимальным. Этим мы обосновали следующее байесовское правило, обеспечивающее наименьшую вероятность ошибки:

$$\text{принять решение } \begin{cases} w_1, \text{если } p(w_1/x) > p(w_2/x); \\ w_2, \text{в противном случае.} \end{cases}$$

Самим видом решающего правила подчеркивается роль апостериорных вероятностей. Используя уравнение (6.93), можно выразить это правило через условные и априорные вероятности. Заметим, что p(x) в уравнении, с точки зрения принятия решения, роли не играет, являясь всего-навсего масштабным множителем, обеспечивающим равенство $p(w_1/x) + p(w_2/x) = 1$. Исключив его, получим следующее решающее правило, полностью эквивалентное предыдущему:

$$\text{принять решение } \begin{cases} w_1, \text{если } p(x/w_1)p(w_1) > p(x/w_2)p(w_2); \\ w_2, \text{в противном случае.} \end{cases}$$

Чтобы пояснить существо вопроса, рассмотрим крайние случаи. Пусть для

некоторого $\hat{x}$ получено, что = $p(x/w_2)$ так, что конкретное наблюдение не дает информации о состоянии атмосферы. В этом случае решение, которое мы примем, целиком зависят от априорных вероятностей. С другой стороны, если $p(w_1) > p(w_2)$, то состояния природы априорно равновозможны и вопрос о принятии решения опирается исключительно на $p(x/w_j)$ – правдоподобие w_j при данном $\hat{x}$. в общем случае при выборе решения важны обе указанные величины, что и учитывается правилом Байеса, обеспечивающем вероятность ошибки.

Если задачу усложнить, сделав ее более общей, то нужно ввести следующие обобщения:

Допускается использование более одного признака.

Допускается более двух состояний природы.

Допускаются действия, отличные от решения о состоянии природы.

Вводится понятие функции потерь, более общее, чем вероятность ошибки.

Эти обобщения и связанные с ними усложнения формул не должны заменять того обстоятельства, что в основном речь идет о явлениях, по существу таких же, как в рассмотренном простом примере.

Из анализа введенных обобщений легко сделать вывод о том, что параметрический дискриминантный анализ есть частный случай теории статистических решений, рассмотренной в подразд. 3.3, следовательно, для него справедливы все положения этой теории. По этой причине мы не будем подробно останавливаться на методах параметрического дискриминантного анализа отметим только, что в дискриминантном анализе широко используется такое понятие, как расстояние Махаланобиса

$$M = \frac{M[w_2] - M[w_1]}{\delta_1 \delta_2}.$$

Расстояние Махаланобиса позволяет количественно определить по какому признаку легче провести дискриминацию. В числителе представлена разность классов находятся дальше друг от друга и тем разделение лучше. В которые, как отмечалось выше, характеризуют компактность классов. Чем компактность больше (δ меньше), тем легче провести дискриминацию. Таким образом расстояние Махаланобиса показывает возможность разделения классов, чем оно больше. Тем разделимость лучше.

Оценивая с помощью этого критерия различные признаки, можно отобрать те из них, которые позволят осуществить дискриминацию с максимальной эффективностью.

6.4 Компонентный и факторный анализ

6.4.1 Основные понятия и классификация видов компонентного и факторного анализа

При исследовании причинно-следственных связей и разборке прогностических методов у исследователя возникает естественное желание рассмотреть как можно большее количество факторов, влияющих, с его точки зрения, на рассматриваемый процесс. При этом возникает две принципиальные трудности такого рассмотрения: во-первых, проведение многофакторного анализа не только существенно усложняет саму процедуру анализа, но и чрезвычайно затрудняет интерпретацию полученных результатов; во-вторых, возникает проблема «табу» статистического прогноза, для устранения которой необходимо формировать выборки очень больших объемов. В связи с этим одной из важнейших задач статистического анализа метеорологической информации является задача максимального ограничения числа исследуемых факторов без существенного снижения информативности конечных результатов обработки. Для решения данной задачи предложен ряд методов, наибольшее распространение среди которых нашли методы, объединенные в компонентный и факторный анализы. (Методы, используемые в настоящее время в синоптической практике, рассмотрены в разд. 8). Так как цель применения данных видов анализа одна – сокращение размерности факторного пространства, то они будут рассмотрены совместно.

Компонентный анализ (КОА) представляет собой совокупность статистических методов обработки информации, позволяющих сконцентрировать информацию, содержащуюся в исходной совокупности результатов наблюдений, за счет перехода к меньшему числу наиболее информативных факторов (главных компонент).

Сущность КОА состоит в преобразовании исходного факторного пространства в пространство главных компонент, общая дисперсия которых равна общей дисперсии исходной совокупности, но перераспределена таким образом, что большая ее часть приходится на первые компоненты. Благодаря этому появляется возможность перехода к меньшему числу переменных за счет отбрасывания последних компонент, которым соответствуют малые дисперсии.

Факторный анализ (ФА) представляет собой совокупность статистических методов обработки результатов наблюдений, позволяющих сконцентрировать информацию, содержащуюся в исходной совокупности результатов наблюдений, за счет описания большого числа исходных косвенных признаков исследуемого явления через меньшее число более емких с информационной точки зрения внутренних характеристик данного явления.

Сущность методов ФА состоит в переходе от описания исследуемого объекта с помощью большого набора непосредственно измеряемых переменных к описанию с меньшим числом максимально информативных (существенных) переменных, являющихся некоторыми функциями исходных переменных и отражающих наиболее существенные свойства исследуемого объекта.

при проведении каждого из данных видов анализа решаются следующие задачи:

представление исходных данных в виде, удобном для проведения анализа;

выбор вида преобразования исходной совокупности факторов к совокупности факторов меньшей размерности;

осуществление преобразования исходной совокупности факторов;

оценка и интерпретация полученных результатов.

Заметим, что как в КОА, так и ФА не выдвигается никаких предположений относительно свойств результатов наблюдений и, кроме того, они не делятся на результаты и факторы, как это было в предыдущих видах анализа. Единственное условие, которое должно выполняться при этом – однотипность закона распределения результатов наблюдений. Невыполнение данного условия приводит к нарушению аддитивности используемых моделей и неверной интерпретации результатов анализа, которая опирается на изучение свойств исходной совокупности наблюдений.

Предположим, что изучаемое явление описывается набором из n показателей свойств (признаков), каждый из которых наблюдался N раз. Результаты наблюдений сводятся обычно в таблицу (см. табл. 6.8)

Таблица 6.8

Номер наблюдения	Номер признака (показатель свойств)					
	1	2	…	j	…	n
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1n}
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2n}
..	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{in}
..	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
N	y_{N1}	y_{N2}	…	y_{Nj}	…	y_{Nn}

или в матрицу наблюдений

$$Y_{(N,n)} = \|y_{ij}\|_N^n. \quad (6.96)$$

Пояснив физическую сущность, вкладываемую в понятия «наблюдения» и «признаки».

В качестве наблюдений в данном анализе рассматривается множество объектов наблюдений, т. е. множество конкретных исследуемых объектов (барических образований, атмосферных фронтов, систем облачности). Это может быть и множество состояний наблюдаемого объекта, определенных в моменты времени, которым соответствует номер строки.

В качестве признаков рассматриваются показатели свойств, характеризующих объекты наблюдения и непосредственно измеряемые в процессе наблюдения. Так например, циклон может быть охарактеризован значением давления в центре, углом наклона пространственной оси, количеством замкнутых изобар, вертикальной протяженностью и т. д., атмосферный фронт – углом наклона, контрастном температуры, барическими тенденциями, протяженностью облачных полей и т.д.

Рассмотренный информационный массив, представленный матрицей наблюдений (табл. 6.8), включает в себя наблюдения по двум факторам: объекты и признаки. Однако число исследуемых факторов может быть большим и включать в себя такие факторы, как время, условия получения информации, способы наблюдения и др. Таким образом, в общем случае результаты наблюдений можно сгруппировать в форме k-мерного информационного массива (k-мерного куба), где k – число факторов. Однако современные методы компонентного и факторного анализа разработаны только применительно к двумерным информационным массивам. В связи с этим при проведении анализа выбирается конкретный двумерный срез k-мерного информационного массива, который и определяет специфику построения и последующие интерпретации матрицы данных.

Можно предложить ряд факторов, применительно к которым чаще всего проводится анализ, а именно: объекты наблюдения, признаки, время, условия и способы наблюдений.

В настоящее время наиболее распространены исследования с матрицами данных, образованных из трехмерного куба данных: объекты признаки, время. В зависимости от того, какие данные образуют строки и столбцы матрицы данных, принято различать несколько разновидностей анализа, которые представлены в табл. 6.9.

Таблица 6.9

Фактор, соответствующий строкам	Фактор, соответствующий столбцам		
	Объект	Признак	Время
Объект	-	R	T
Признак	Q	-	O
время	S	P	-

Данные разновидности обозначены символами R, T, Q, O, S, P и называются R-анализом, T-анализом, Q-анализом и т. д. Особенности каждого из видов состоят в следующем.

В R-анализе изучается изменение признаков от объекта к объекту. При этом определяются группы признаков со сходным характером вариаций, которые и являются основой для формирования главных компонент в КОА или факторов в ФА. Эти компоненты и факторы рассматриваются как обобщенные характеристика, содержащие информацию, имеющуюся в исходном наборе признаков.

В Q-анализе, наоборот, выявляются группы объектов, имеющих сходные свойства по большому набору признаков, т. е. решается, по существу, задача классификации объектов, при которой выявленные компоненты (факторы) рассматриваются как критерии для классификации.

В T-анализе признак фиксирован, а выявляются периоды времени с характерными для каждого периода распределением значений признака по объектам наблюдения.

В S-анализе признак также фиксирован, а каждый элемент матрицы данных является значением некоторого (одного или того же для всех столбцов) признака на j-том объекте в момент времени t. Такая постановка позволяет изучать динамику объектов наблюдения, т. е. выявлять группы объектов со сходным типом изменений во времени.

В O-анализе выявляются периоды времени со сходными (внутри периода) сочетанием значений признаков. Выявленные компоненты (факторы) трактуются при этом как соответствующие периоды.

В P-анализе исследуется характер вариации признаков во времени и отыскиваются факторы (компоненты) как некоторые обобщенные параметры, хорошо описывающие эту вариацию.

Аналогичным образом могут быть образованы другие виды анализа по другим факторам. Замечательной особенностью как компонентного, так и факторного анализа является то, что любой из перечисленных выше видов анализа проводится с помощью одних и тех же методов. Специфика отдельных видов

анализа состоит в интерпретации полученных результатов.

В связи с этим рассмотрение данных методов проведем на примере R-анализа, тем более что по сравнению с другими описанными выше видами он является наиболее полно разработанным.

В заключение сделаем два важных для практической работы замечания:

1. Модели КОА и ФА опираются на допущение о равноважности всех объектов наблюдения. Если это условие не выполняется, то объектами должны быть приписаны соответствующие коэффициенты важности.

2. Весьма ответственный является процедура формирования матрицы данных. Корректный выбор объектов наблюдения и признаков, включаемых в матрицу данных, полностью определяет корректность результатов, полученных при анализе.

6.4.2 Компонентный анализ

Исходными данными для компонентного анализа являются результаты наблюдений, представленные в виде табл. 6.8 или матрицы наблюдений (6.96).

Матрицу наблюдений можно интерпретировать как реализацию n-мерного случайного вектора

$$\hat{Y}_{<n>} = \langle \hat{y}_1, \hat{y}_2, \dots, \hat{y}_n \rangle \qquad (6.97)$$

в результате N наблюдений.

Как известно, свойства данного случайного вектора с достаточной для практики точности описываются вектором математических ожиданий

$$\bar{Y}_{<n>} = \langle \bar{y}_1, \bar{y}_2, \dots, \bar{y}_n \rangle \qquad (6.98)$$

и корреляционной матрицей

$$K_{(n)} = \left\| K_{ij} \right\|_N^n, \qquad (6.99)$$

где $K_{ij} = \mathrm{M}\left[(\hat{y}_i - \bar{y}_i)(\hat{y}_j - \bar{y}_j)\right]$, или матрицей коэффициентов корреляций

$$R_{(n)} = \left\| r_{ij} \right\|_N^n, \qquad (6.100)$$

где

$r_{ij} = \frac{K_{ij}}{\delta_i \delta_j}$.

Ранее было показано, что матрицы (6.99) и (6.100) могут быть построены на основе различных видов корреляций. В компонентном и факторном анализе чаще всего используются полные (парные) корреляции и коэффициенты корреляций, хотя в принципе допустимо использование и других видов корреляций.

Необходимо заметить, что результаты наблюдений, относящиеся к различным признакам, чаще всего имеют различную размерность, поэтому перед вычислением корреляционной матрицы эти результаты необходимо привести к

стандартной или нормированной форме. Нормировка осуществляется традиционным способом, т. е. вместо y_{ij} используются переменные z_{ij}, полученные на основе выражения

$$z_{ij} = \frac{y_{ij} - \bar{y}_j}{\delta_j}$$

Для простоты изложения предположим, что признаки имеют одинаковую размерность.

Модель компонентного анализа основывается на предположении, что любой j-й признак $\hat{y}_j$ может быть представлен в виде линейной комбинации главных компонент $\hat{f}_i$:

$$\hat{y}_j = a_{1j}\hat{f}_1 + a_{2j}\hat{f}_2 + \cdots + a_{nj}\hat{f}, \qquad j = \overline{1,n}, \tag{6.101}$$

где f_1, – главные компоненты;

$f_2, \ldots, f_n$ – вес i-ой главной компоненты в j-ой переменной.

a_{ij}

Под главными компонентами $\hat{f}_i$ понимаются некоррелированные между собой безмерные переменные, представляющие некоторую линейную комбинацию n переменных:

$$\hat{f}_i = a_{1i}\hat{y}_1 + a_{2i}\hat{y}_2 + \cdots + a_{in}\hat{y}_n, \tag{6.102}$$

Главные компоненты определяются таким образом, чтобы первая из них давала максимально возможный вклад в суммарную дисперсию результатов наблюдений, вторая – максимальный вклад в дисперсию, оставшуюся после исключения дисперсии, соответствующей первой главной компоненте, и т.д. в этом случае задача анализа главных компонент сводится к следующему: найти такое линейное ортогональное преобразование n наблюдаемых признаков, чтобы получить совокупность n некоррелированных нормированных переменных f, дисперсии которых обладают свойствами:

$$\delta^2[\hat{f}_1] \geq \delta^2[\hat{f}_2] \geq \ldots \geq \delta^2[\hat{f}_n]. \tag{6.103}$$

Такое преобразование эквивалентно преобразованию исходной корреляционной матрицы К к матрице вида

$$F^T F = \begin{vmatrix} \delta^2[\hat{f}_1] & 0 & 0 & 0 \\ 0 & \delta^2[\hat{f}_2] & 0 & 0 \\ 0 & 0 & 0 & \delta^2[\hat{f}_n] \end{vmatrix} \tag{6.104}$$

Для данной матрицы выполняется условие

$$\sum_{i=1}^{n} \delta^2[\hat{f}_i] = \sum_{j=1}^{n} \delta^2[\hat{y}_j]. \tag{6.105}$$

Зная дисперсию, соответствующую каждой главной компоненте, можно оценить вклад этой компоненты в формирование общей дисперсии совокупности

результатов наблюдений и тем самым выделить существенные компоненты, которые целесообразно включить в дальнейший анализ. Практика показывает, что число таких компонент составляет 10 – 25% от общего числа компонент, что и позволяет существенно сократить размерность совокупности признаков.

Таким образом, в компонентном анализе осуществляется описание исходного вектора $\hat{Y}_{<n>} = \langle \hat{y}_1, \hat{y}_2, \dots, \hat{y}_n \rangle >$ n линейными комбинациями переменных y_1, y_2, …, y_n (главными компонентами) так, что каждая линейная комбинация содержала как можно большую часть вариации по $\hat{y}$ и, в то же время, была линейно независима от всех других главных компонент.

Поясним сказанное следующим примером. Предположим, что имеется две переменные $\hat{y}_1$ и $\hat{y}_2$, представляющие собой нормально распределенные коррелированные случайные величины с математическими ожиданиями $\bar{y}_1$ и $\bar{y}_2$ и корреляционной матрицей $К_{(2)}$. При положительной корреляции этим переменным соответствует эллипс рассеивания, показанный на рис. 6.18. при представлении системы переменных $\langle \hat{y}_1, \hat{y}_2 \rangle$ в форме главных компонент первая главная компонента будет соответствовать большой оси эллипса (прямая I), а вторая главная компонента – малой оси эллипса (прямая II). В общем случае, когда имеется n переменных (признаков), каждая компонента определяет одну из осей n-мерного эллипсоида.

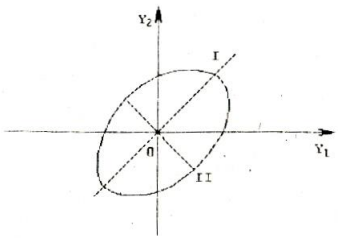


Рис. 6.18

Перепишем выражения (6.101) и (6.102) в векторно-матричной форме. Для этого введем:

вектор результата наблюдения

$$\hat{Y}_{<n>} = \langle \hat{y}_j \rangle_n = \langle \hat{y}_1, \hat{y}_2, \dots, \hat{y}_n \rangle; \qquad (6.106)$$

вектор главных компонент

$$\hat{F}_{<n>} = \langle \hat{f}_i \rangle_n = \langle \hat{f}_1, \hat{f}_2, \dots, \hat{f}_n \rangle; \qquad (6.107)$$

матрицу ортогонального преобразования A, определяемую следующим образом:

$$A = \left\| a_{ij} \right\|_n^n = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1i} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2i} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{j1} & a_{j2} & \dots & a_{ji} & \dots & a_{jn} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{ni} & \dots & a_{nn} \end{vmatrix} \quad (6.108)$$

Тогда выражение (6.101) запишется в виде

$$\hat{Y}_{<n>} = \hat{F}_{<n>} A^T \quad (6.109)$$

Так как матрица А является ортогональной, т. е.

$A^T A = I$,

где I – единичная матрица, то следовательно, выражение (6.107) можно представить в виде

$\hat{F}_{<n>} = \hat{Y}_{<n>} A$.

Если матрица А задана, то дисперсию случайного вектора $\hat{F}_{<n>}$ можно определить следующим образом:

$$D\left[\hat{F}_{<n>}\right] = D\left[\hat{Y}A\right] = A^T D\left[\hat{Y}_{<n>}\right] A = A^T K A \quad (6.110)$$

Как уже отмечалось ранее, в компонентном анализе первая главная компонента выбирается таким образом, чтобы она имела максимальную дисперсию. В общем случае эта задача не имеет однозначного решения.

Для преодоления этой трудности вводится условие нормировки для матрицы А, состоящее в том, чтобы

$$A^T A = 1 \quad (6.111)$$

т.е. в требовании, чтобы эта матрица имела единичную длину. Тогда задача определения первой главной компоненты может быть сформулировано следующим образом: максимизировать выражение

$$A^T \, KA \quad (6.112)$$

при условии (6.111).

данная задача представляет собой задачу максимизации линейной формы (6.112) при ограничении (6.111). ее решение состоит в следующем.

Используя метод множителей Лагранжа, образуем функцию Лагранжа

$\Psi = A^T \, KA - \lambda(A^T A - 1)$,

где λ – множитель Лагранжа.

Вектор частных производных имеет вид

$\frac{\partial \Psi}{\partial A} = 2KA - 2\lambda A$.

После приравнивания его к нулю получим уравнение

$$(K - \lambda I)A = 0 \quad (6.113)$$

которое представляет собой не что иное, как характеристическое уравнение. Оно имеет решения, отличные от нуля, если определитель матрицы (К – λI)

равен нулю, т. е. матрица (К – λI) должна быть вырожденной. Определитель данного уравнения

$$|К - \lambda I| = 0 \qquad (6.114)$$

представляет собой алгебраическое уравнение относительно λ. Таким образом, для решения данного уравнения необходимо найти n характеристических корней матрицы К. обозначим эти корни в виде $\lambda_1 \geq \lambda_2 \geq \lambda_n$. Для того, чтобы определить, какой из этих корней можно использовать для нахождения характеристического вектора, минимизирующего (6.112), умножим (6.113) слева на A^T. Тогда. Учитывая, что $A^T A = 1$ получим

$A^T(К - \lambda I)A = A^T КА = \lambda = 0$

или

$$A^T КА = \lambda. \qquad (6.115)$$

Однако из (6.110) видно, что А К А является дисперсией вектора $\hat{F}_{<n>}$. Таким образом, чтобы максимизировать дисперсию, необходимо выбрать небольшой характеристический корень корреляционной матрицы К. обозначим этот корень через λ_1, а нормированный характеристический вектор, соответствующий данному корню, через A_1. Тогда первая главная компонента задается уравнением

$$F_{<n>1} = Y_{<n>}A_1 \qquad (6.116)$$

с дисперсией, равной λ_1.

Аналогичным образом вторая главная компонента с дисперсией λ_2 определяется в виде

$F_{<n>2} = Y_{<n>}A_2$

где A_2 - нормированный характеристический вектор, соответствующий характеристическому корню λ_2, матрицы К, и т. д.

Если указанные вычисления провести для рассмотренного ранее примера, то получим, что первая главная компонента соответствует большой оси эллипса, причем половина длины этой оси пропорциональна $\sqrt{\lambda_1}$, а вторая главная компонента – малой полуоси эллипса, и половина ее длины пропорциональна $\sqrt{\lambda_2}$.

Таким образом, вычисление главных компонент совокупность результатов наблюдений сводится к анализу ее матрицы корреляционных моментов и вычислению собственных чисел и собственных векторов этой матрицы. Преобразование к главным компонентам не меняет сумму дисперсий, а только перераспределяет ее таким образом, чтобы на первые компоненты приходилась большая часть дисперсии. При этом первая главная компонента $F_{<n>1}$ представляет собой линейную комбинацию n переменных с коэффициентами, равными нормированному характеристическому вектору, соответствующему наибольшему характе-

ристическому значению матрицы К. Вторая главная компонента $F_{<n>2}$представляет собой линейную комбинацию n переменных с коэффициентами, равными нормированному характеристическому вектору, соответствующему второму по величине характеристическому значению матрицы К, и т. д., вплоть до n-ой главной компоненты $F_{<n>n}$. Дисперсия каждой главной компоненты равна соответствующему собственному значению, и все компоненты линейно не зависят друг от друга. Аким образом, каждая компонента просто определяет одну из осей n-мерного эллипсоида рассеивания. Грубо говоря, в результате компонентного анализа происходит преобразование системы координат таким образом, что главные компоненты становятся координатными осями.

Определение главных компонент – первый этап компонентного анализа. На последующих этапах решается ряд последовательных задач, конечным итогом которых является определение ограниченного состава компонент, которые целесообразно включать в дальнейшую обработку. К числу этих задач относятся:

задача оценивания существенности компонент;

задача оценивания достоверности полученных результатов;

задача выделения существенных компонент.

Первая из рассмотренных задач сводится к оцениванию «важности» i-ой главной компоненты в описании совокупности результатов наблюдений. В качестве показателя существенности (важности) i-ой компоненты целесообразно выбрать отношение

$$\beta_i = \frac{\lambda_i}{\sum_n^i \lambda_i}. \qquad (6.117)$$

Очевидно, что чем больше β_i по сравнению с другими β_j(j≠i), тем более существенный вклад вносит i-я главная компонента в общую дисперсию результатов наблюдений.

Необходимость решения второй задачи обусловлена тем, что главные компоненты определяются на основе случайной совокупности результатов наблюдений и, следовательно, полученные величины являются оценками собственных значений и собственных векторов. В связи с этим оценка достоверности полученных результатов сводится к проверке гипотез, касающихся собственных значений и собственных векторов. При этом можно выдвинуть два типа нулевых гипотез.

Первая из них опирается на предположение, что все собственные значения равны, т. е.

H_0: $\lambda_1 = \ldots = \lambda_n$,

а альтернативная гипотеза имеет вид:

H_1: $\lambda_1 \neq \lambda_2 \neq \ldots \neq \lambda_n$.

В качестве показателя согласованности при проверке гипотезы H_0 можно

использовать различные статистики, в частности основанную на отношении правдоподобия, имеющую вид

$$U_1 \left| \frac{|K|}{\left[\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right]^n} \right|^{(N-1)/2} \tag{6.118}$$

или более удобную с практической точки зрения статистику

$$U_2 = -2lnU_1. \tag{6.119}$$

Для больших выборок распределение статистики U_2 приближается к распределению

$$\chi^2 = -(n-1)\left[ln|K| - n\ ln\frac{\sum_{i=1}^{n}\lambda_i}{n}\right] \tag{6.120}$$

с ½(n+1)n – 1 степенями свободы.

Учитывая свойства собственных значений, выражение (6.120) можно переписать в виде

$$\chi^2 = -(n-1)\left[\sum_{i=1}^{n} ln\lambda_i - n\ ln\frac{\sum_{i=1}^{n}\lambda_i}{n}\right] \tag{6.121}$$

Для корреляционной матрицы R показатель U_1 принимает вид

$U_1 = |R|^{(N-1)/2}$,

а закон распределения

$\chi^2 = -(n-1)\ ln|K|$

с ½(n –1) степенями свободы.

Критическая область является двусторонней. Проверка гипотезы проводится описанным в гл. ; методом.

На практике часто проверяется гипотеза следующего вида. Предположим, что первые m главных компонент $f_1, f_2, \ldots, f_m$ велики и их дисперсии составляют основную часть, например, 90% общей дисперсии. При этом по отношению к оставшимся n-m компонентам целесообразно проверить гипотезу о их существенном различии. Если эта гипотеза подтверждается, то использование этих компонент становится ненужным. Таким образом проверяется гипотеза

H_0: $\lambda_{m+1} = \lambda_{m+2} = \ldots = \lambda_n$,

И альтернативная к ней гипотеза H_1: собственные значения не равны.

Показатель согласованности для проверки гипотезы H_0 основывается на отношении правдоподобия и имеет вид

$$U_3 = \left[\frac{\sum_{i=1}^{n}\lambda_i}{\lambda^{(n-m)}\sum_{i=1}^{m}\lambda_i}\right]^{(N-1)/2}, \tag{6.122}$$

А его закон распределения приближается для больших выборок к распределению *хи*-квадрат

$\chi^2 = -(n-1)\left[\sum_{i=m+1}^{n}\lambda_i - (n-m)\, ln\frac{\sum_{i=m+1}^{n}\lambda_i}{n-m}\right]$

с ½(n –m)(n-m+1) степенями свободы.

Последняя из перечисленных задач – задача выделения существенных компонент – решается на основе показателя важности β_i (6.117). для ее решения устанавливают величину обобщенной дисперсии, которую должны описывать существенные компоненты. В зависимости от задачи исследования величина этой дисперсии может составлять от 80 до 95%. Затем определяется число первых компонент, которые обеспечивают данную величину дисперсии, т. е.

$$\beta_1 + \beta_2 + \cdots + \beta_m \geq 0{,}8 \div 0{,}95$$

Практика проведения компонентного анализа показывает, что число существенных компонент составляет 10 – 25% от общего числа признаков. Таким образом, метод главных компонент позволяет значительно сократить число исследуемых признаков.

Наиболее ответственным заключительным этапом анализа главных компонент является этап интерпретации выделенных существенных компонент. Сложность данного этапа определяется тем, что главные компоненты представляют собой линейную комбинацию исходных признаков и в явном виде не поддаются содержательной интерпретации, за исключением отдельных частных случаев.

Так как получаемое решение является однозначным, то какие-либо дальнейшие процедуры преобразования главных компонент, с целью более четкого определения, невозможны, что является существенным недостатком компонентного анализа. Попытки интерпретировать главные компоненты как некоторые скрытые (неявные) обобщенные параметры исследуемого объекта не имеют под собой достаточного научного обоснования, т. е. анализ главных компонент целесообразно рассматривать не более как метод определения осей эллипсоида. Он может быть методом обнаружения обобщенных факторов, скрытых в исходных данных, если физический смысл задачи допускает их четкую интерпретацию, а может и не быть им.

6.4.3. Факторный анализ

Исходные данные для факторного анализа аналогичны исходным данным в компонентном анализе, т. е. включают в себя матрицу наблюдений (6.96), вектор математических ожиданий (6.98) и корреляционную матрицу (6.99). При этом на практике чаще всего используются эти данные в нормативном виде, т. е. в виде матрицы наблюдений

$$Z_{(N,n)} = \left\|z_{ij}\right\|_N^n, \tag{6.123}$$

и корреляционной матрицы

$$R = \left\|z_{ij}\right\|_N^n, \tag{6.124}$$

где

$z_{ij}\frac{\bar{y}_{ij}-y_j}{\delta_j}$,

$r_{ij} = \mathrm{M}\left[\hat{z}_i\hat{z}_j\right]$.

Однако модель факторного анализа существенно отличается от модели компонентного анализа. Это отличие состоит во введении следующих предположений.

1. Предполагается, что любой j-ый признак может быть представлен в виде линейной комбинации факторов ψ_i(i=$\overline{1,r}$), число которых существенно меньше числа исходных признаков, т. е.

$$\hat{z}_j = b_{j1}\psi_1 + b_{j2}\psi_2 + \cdots + b_{jr}\psi_r \qquad (6.125)$$

где $\psi_1, \psi_2, \ldots, \psi_r$ – факторы;

$b_{j1}, b_{j2}, \ldots, b_{jr}$ – коэффициенты, характеризующие влияние каждого из факторов на j-ый признак и называемые факторными нагрузками.

Под фактором ψ_i(i=$\overline{1,r}$) понимается некоторый неизвестный и не включенный в исследуемую совокупность признак, описывающий влияние на исходные признаки.

2. предполагается, что факторы ψ_i(i=$\overline{1,r}$) по характеру их влияния на признаки z_j(j=$\overline{1,n}$) могут быть разделены на две группы: общие, которые в дальнейшем будем обозначать f_i, причем i=$\overline{1m}$, и m = r – 1; характерные, которые обозначаются символом V_j(j=$\overline{1,n}$). С учетом этого зависимость признаков от факторов может быть представлена в виде

$$\hat{z}_j = a_{j1}f_1 + a_{j2}f_2 + \cdots + a_{jm}f_m + d_jV_j, \qquad (6.126)$$

где $a_{j1}, a_{j2}, \ldots, a_{jm}$ – факторные нагрузки $b_{j1}, b_{j2}, \ldots, b_{jm}$ при общих факторах;

d_j – факторная нагрузка b_{jr} при характерном факторе V_j.

Под общим фактором понимается фактор ψ_i, оказывающий влияние более чем на один признак z_j исходной совокупности признаков.

Таким образом, в ФА предполагается, что можно определить совокупность расчетных признаков, оказывающих влияние на ряд исходных признаков (совокупность общих факторов), и совокупность расчетных признаков, оказывающих влияние только на один из исходных признаков (совокупность характерных факторов).

Целью факторного анализа является определение такой совокупности общих факторов, которая наилучшим образом описывала бы совокупность результатов наблюдений при условии, что число данных факторов должно быть существенно меньше числа исходных признаков.

Показателем, характеризующим качество описания совокупности результатов наблюдений с помощью ограниченного числа факторов, как и в компонентном анализе, является дисперсия.

Определим ее в предположении, что признаки представимы в виде (6.126)

$\delta^2_{\hat{z}_j} = M\left[\left(\hat{z}_j - \bar{z}_j\right)^2\right]$, где j = $\overline{1,n}$

Так как $\bar{z}_j = 0$ по условию, то получим

$$\delta^2_{\hat{z}_j} = M\left[\hat{z}_j^2\right] = M\left[\left(a_{j1}\hat{f}_1 + a_{j2}\hat{f}_2 + \cdots + a_{jm}\hat{f}_m + d_j\hat{V}_j\right)^2\right] =$$
$$= \sum_{i=1}^{m} a_{ij}^2 M\left[\hat{f}_j^2\right] + d_j^2 M\left[\hat{V}_j^2\right] + 2\sum_{k<q=1}^{m} a_{jk}a_{jq}\, r_{\hat{f}_k\hat{f}_q} + \qquad (6.127)$$
$$2d_j \sum_{i=1}^{m} a_{ji} r_{\hat{f}_i\hat{V}_j}.$$

Факторный анализ ожжет проводится при нескольких предположениях относительно свойств общих и характерных факторов.

Прежде всего, всегда предполагается, что характерные факторы попарно некоррелированы ни между собой, ни с одним из общих факторов.

Относительно общих факторов может быть выдвинуто одно из двух предположений: предположение о том, что они попарно некоррелированы, и предположение о том, что они коррелированны между собой. Факторный анализ, выполняемый при первом предположении, называется анализом с некоррелированными или отрицательными факторами. Факторный анализ, выполненный при втором предположении, называется анализом с коррелированными (не отрицательными или косоугольными факторами). На практике наибольшее распространение получил ФА с ортогональными факторами, который и будет здесь рассмотрен.

Кроме того, на практике факторы чаще всего рассматриваются в нормированном виде.

С учетом последнего обстоятельства и ортогональности факторов две последние суммы в выражении (6.127) будут равны нулю и, следовательно, получаем

$$\delta^2_{\hat{z}_j} = \sum_{i=1}^{m} a_{ij}^2 + d_j^2, \qquad (6.128)$$

так как $M\left[\hat{f}_j^2\right] = 1$ и $M\left[\hat{V}_j^2\right] = 1$.

По условию $\delta_{\hat{z}_j}^2 = 1$, поэтому выражение (6.128) можно записать в виде

$$\delta_{\hat{z}_j}^2 = \sum_{i=j}^{m} a_{ij}^2 + d_{ji}^2 = 1. \quad (6.129)$$

Это выражение определяет разложение общей дисперсии по дисперсиям факторов, при этом квадраты факторных нагрузок a_{ij}^2(i = $\overline{1m}$,j = $\overline{1,n}$) характеризуют вклад i-го фактора, а квадрат нагрузки d_j - вклад j-го характерного фактора в общую дисперсию.

Вкладом фактора $\hat{f}_i$ в дисперсию признака z_i называется квадрат соответствующей факторной нагрузки a_{ij}^2.

Суммарным вкладом фактора $\hat{f}_i$ в дисперсию всех признаков, или просто вкладом фактора $\hat{f}_i$, называется число, определяемое выражением

$$\varepsilon_i = \sum_{j=1}^{m} a_{ij}^2 \quad (6.130)$$

Как уже отмечалось, целью ФА является наиболее полное объяснение дисперсии исходных признаков с помощью общих факторов, так как в этом случае достигается полная замена исходных признаков выделенными факторами. Следовательно, фактора должны выбираться таким образом, чтобы увеличивать ту часть дисперсии признака $\hat{z}_j$, которую определяют нагрузки общих факторов. Поэтому данные нагрузки объединяют в одно целое, называемое общностью и обозначают символом h_j^2, т. е.

$$h_j^2 = \sum_{i=1}^{m} a_{ji}^2$$

Общность служит для определения вклада всех факторов и каждого в отдельности и объяснение дисперсии признаков. Втрое слагаемое в выражении (6.129) называют характерностью. Она представляет ту часть дисперсии j-го признака, которую не удается объяснить общими факторами, поэтому ее вклад целесообразно уменьшать.

Учитывая сказанное, выражение (6.129) можно записать в виде

$$\delta_{\hat{z}_j}^2 = h_j^2 + d_j^2 = 1. \quad (6.131)$$

Задачей факторного анализа, следовательно, является определение числа и

состава общих факторов, наилучшим образом представляющих свойства исходной совокупности результатов наблюдений. В заключение сделаем ряд замечаний, относящихся к сравнению КОА и ФА, и отметим вытекающие отсюда особенности проведения факторного анализа.

1. из сравнения моделей (6.101) и (6.126) видно, что если в компонентном анализе число главных компонент равно числу исходных признаков, то в факторном анализе число факторов с самого начала выбирается существенно меньшим. Так, в частности, количество факторов m и количество исходных признаков n обычно выбирается в соотношении

$$(\mathrm{n} + \mathrm{m}) < (n - m)^2. \qquad (6.132)$$

2. в отличие от КОА совокупность общих факторов определяется неоднозначно, что допускает ее преобразование к виду, обеспечивающему наилучшую содержательную интерпретацию факторов.

3. В связи с тем, что в ФА используются характерные факторы, влияние которых на результаты анализа желательно исключить, и, кроме того, число факторов меньше числа исходных признаков, в ФА вместо исходной матрицы R используются специальным образом построения на ее основе редуцированная матрица коэффициентов корреляции R_0.

Рассмотренные выше основные положения ФА позволяют представить процедуру его проведения в виде схемы, изображенной на рис. 6.18. Процедура ФА включает в себя ряд этапов, а именно:

На первом этапе по исходным данным Z вычисляется матрица коэффициентов корреляции и на ее основе редуцированная корреляционная матрица R_0.

На втором этапе с помощью матрицы R_0 определяется факторная матрица, т. е. определяется число m факторов и факторные признаки.

На третьем этапе осуществляется преобразование (вращение) факторов с целью их содержательной интерпретации. Заметим, что в отличие от КОА факторный анализ позволяет в определенной степени выявлять скрытые обобщенные признаки исследуемого объекта.

На четвертом этапе осуществляется оценивание значений факторов.

Для нахождения методов определения факторов перепишем выражение (6.125) и (6.126) в векторно-матричной форме. Для этого введем следующие векторы и матрицы:

матрицы данных

$\mathrm{z} = \langle z_j \rangle_n = \left\| z_{ij} \right\|_N^n$,

где z_j – j–ый вектор-столбец матрицы данных;

матрицу факторов

$\psi = \langle \psi_1, \psi_2, \dots, \psi_r \rangle$,

где ψ_p(p = $\overline{1,r}$) – вектор значений p-го фактора на объектах наблюдений;

матрицу факторных нагрузок

B = $\|b_{ik}\|_n^r$;

матрицу общих факторов

F = $\langle F_1, \dots, F_m \rangle$,

где F_i = (i = $\overline{1m}$) – вектор значений i-го общего фактора на объектах наблюдения;

матрицу факторных нагрузок общих факторов

A = $\|a_{ij}\|_n^m$;

Диагональную матрицу факторных нагрузок характерных факторов

D = $\|d_{jj}\|_n^n = \langle D_1, \dots, D_n \rangle$,

где $D_j = < 0,0, \dots ,0, d_j, 0, \dots ,0 >^T$,

вектора нагрузок j-го характерного фактора.

С учетом данных обозначений выражение (6.125) запишется в виде

$$Z^T = B\psi^T, \tag{6.133}$$

а выражение (6.126) – в виде

$$Z^T \text{A} \cdot F^T + D \cdot V^T. \tag{6.134}$$

Заметим, что в литературе очень часто запись данных выражение дается по отношению не к признакам, а по отношению к исходной матрице наблюдений, что не вполне соответствует моделям (6.125) и (6.126).

Методы определения факторов опираются на два фундаментальных положения: соотношение для определения матрицы коэффициентов корреляции R и так называемую фундаментальную теорему ФА.

Соотношение для матрицы коэффициентов корреляции определяется выражением (6.124), которое в матричной форме записи имеет вид

$$R = Z^T \cdot Z. \tag{6.135}$$

Фундаментальная теорема ФА применительно к введенным выше обозначениям имеет следующую формулировку.

Теорема 1. Если результаты наблюдений представимы в виде линейной комбинации r факторов, т. е.

$Z^T = B \cdot \psi^T$,

то матрица коэффициентов корреляции может быть выражена с помощью матрицы факторных нагрузок в виде

$$R_{(n)} = Br_{(r)}B^T, \tag{6.136}$$

если факторы коррелированны и матрица коэффициентов корреляции факторов $R = \psi^T \cdot \psi$, или в виде

$$R = \text{BB}^T, \tag{6.137}$$

если факторы некоррелированы (ортогональны).

Доказательство. На основе (6.135) получим

$$R_{(n)} = Z^T Z = B\psi^T [B\psi^T]^T = B\psi^T \psi B^T.$$

Произведение $\psi^T \psi$ есть не что иное, как матрица коэффициентов корреляции $R_{(r)}$, откуда получаем справедливость (6.136).

При некоррелированности факторов $R_{(r)} = I$, откуда получаем

$$R_{(n)} = B\, I\, B^T = B\, B^T,$$

т. е. справедливость равенства (6.137).

применяя фундаментальную теорему к моделям (6.126) или (6.134), получают различные методы определения факторов. В связи с тем, что ФА приводится с помощью ЭВМ и для каждой ЭВМ имеется свое описание библиотеки стандартных программ, мы не будем останавливаться на описании методов.

После определения факторов производится их анализ, который включает в себя решение ряда задач, основными из которых являются задача оценивания качества полученного решения и задача содержательной интерпретации факторов. Для осуществления последней, как уже отмечалось выше, часто возникает необходимость преобразования (вращения) факторов, только после выполнения которого они могут считаться окончательно определенными и, следовательно, по отношению к ним целесообразно решить задачу оценивания качества полученного решения.

Сущность задачи вращения сводится к следующему. Необходимо вращать факторы таким образом, чтобы как можно больше коэффициентов факторных нагрузок оказывались нулевыми и, следовательно, наблюдаемые признаки зависели бы от возможно меньшего числа факторов, а также чтобы не нулевые нагрузки оказывались максимально возможными. Для наглядности этим положениям можно дать следующую геометрическую интерпретации. (рис. 6.18).

Предположим. Что в результате анализа получено два общих фактора f_1 и f_2 для трех исходных признаков z_1, z_2, z_3. На основе формулы (6.126) факторы с геометрической точки зрения можно представить в виде ортогональных осей, образующих факторное пространство, а признаки – в виде векторов в данном пространстве. На рис. 6.19а показан случай, когда факторы получены без учета описанных выше требований. Так как факторы допускают вращение (при сохранении ортогональности), то поворот их в положение изображенное на рис. 6.19б, как это видно в большей степени соответствует этим требованиям. Такое расположение облегчает и интерпретацию факторов, выделяя признаки в наибольшей степени связанные с каждым из факторов.

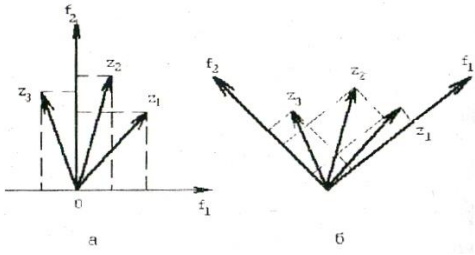


Рис. 6.19

В настоящее время существует две группы методов решения задачи вращения: геометрические и алгебраические. Первые обладают большей наглядностью, но применимы только для трехмерных пространств. Вторые имеют меньшую наглядность, но применимы для любого числа факторов. Среди данных методов наибольшее распространение получил метод вращения, основанный на так называемом веримакс-критерии, удовлетворяющем рассмотренным выше требованиям.

Сущность данного метода состоит в следующем. Наиболее экономное описание точки в двумерном пространстве соответствует случаю, когда ось координат проходит через эту точку. Таким образом, вращение координат необходимо осуществлять так, чтобы стремиться к этому идеальному случаю. Очевидно, что если одна из осей координат приближается к точке, то произведение координат точки начинает уменьшаться. Следовательно, в качестве показателя, характеризующего положение координат, целесообразно выбрать показатель

$$U = \sum_{k<i=1}^{m} \sum_{i=1}^{N} (a_{ik} a_{i1})^2, \qquad (6.138)$$

а критерием оптимального расположения осей является такой выбор положения факторов, при котором показатель U принимает минимальное значение.

Веримакс-критерий использует данный принцип, но имеет более сложный вид. Для его описания введем следующие обозначения

$A = \|a_{ij}\|$ – исходная ортогональная факторная матрица;

$B_1 = \|b_{ij}\|$ - финальная (окончательная) факторная матрица, полученная в результате ортогонального вращения;

T – ортогональная матрица преобразования.

Следовательно,

$$B_1 = AT. \qquad (6.139)$$

Учитывая (6.139), получаем

$$\sum_{j=1}^{N} b_{1ij}^2 = \sum_{j=1}^{m} a_{ij}^2 = h_i^2, \quad (i=\overline{1,m}). \qquad (6.140)$$

Веримакс-критерий формируется следующим образом: факторная матрица

удовлетворяет сформулированным выше требованиям, если ее нагрузки обеспечивают максимум функции

$$U = m\sum_{j=1}^{m}\sum_{i=1}^{N}\left(\frac{b_{ij}}{h_i}\right)^4 - \sum_{j=1}^{m}\left[\sum_{i=1}^{N}\frac{b_{ij}}{h_i^2}\right]^2. \quad (6.141)$$

Таким образом, простота интерпретации связывается с дисперсией нормированных квадратов факторных нагрузок. Если же эта дисперсия максимальна, то фактор будет наиболее интерпретируемым, так как его нагрузки будут близки или к единице, или к нулю.

Сущность оценивания качества полученного решения состоит, как и в методе главных компонент, в проверке ряда гипотез относительно полученных факторов и факторных матриц, так как их элементы получены на основе случайной совокупности результатов наблюдений. Однако из-за сложности проведения такой проверки на практике ограничиваются определением точности полученных оценок факторов. В качестве показателя, являющегося мерой оценки точности фактора, применяется коэффициент множественной корреляции.

В заключение заметим, что требование ортогональности факторов является весьма сильным ограничением, существенно усложняющим процедуру интерпретации, т. е. ортогональные факторы не всегда позволяют достаточно полно описать признаки, рассредоточенные в факторном пространстве. В связи с этим в последние годы проводятся интенсивные исследования по разработке методов ФА, в которых данное ограничение отсутствует, так называемого анализа с косоугольными факторами. Однако широко практического распространения эти методы пока не получили.

Результаты использования факторного анализа для отбора информативных прогностических признаков, полученные А. В. Чаплыгиным, свидетельствует о его большей эффективности по сравнению со стандартными процедурами отбора предикторов, используемыми в синоптической практике (гл. 8).

7 СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

7.1 Предварительные замечания

Одним из основных методов исследования закономерностей атмосферных процессов является метод математического моделирования. Особенно большое значение этот метод имеет при анализе крупномасштабных процессов (циклонической деятельности, общей циркуляции атмосферы и др.). Дело в том, что при изучении таких объектов, с одной стороны, недопустимы натурные эксперименты (например, эксперименты, направленные на изменение климата в результате уничтожения арктических ледяных полей, перекрытия морских проливов), поскольку для части регионов и государств неизбежны негативные (к тому же трудно предсказуемые) последствия соответствующих мероприятий.

С другой стороны, при создании технических моделей крупномасштабных метеорологических объектов не удается обеспечить подобие рассматриваемых полей, и, таким образом, построение и исследование математических моделей крупномасштабных атмосферных процессов является, по крайней мере в настоящее время, единственным методом их изучения.

Принято различать два типа математических моделей: аналитические и имитационные модели.

При построении аналитических моделей метеорологических процессов

используется система уравнений гидротермодинамики (отсюда название моделей – динамические). Система включает дифференциальные уравнения в частных производных, и ее аналитическое решение возможно лишь при серьезных упрощающих предположениях, (например, стационарности процессов, отсутствии вязких напряжений и др.) поэтому для решения систем «полных» (неупрощенных) уравнений динамики атмосферы используются различные численные методы.

Одним из численных методов исследования аналитических моделей является рассматриваемый в подразд. 7.2 метод статистических испытаний (метод Монте-Карло).

Имитационные модели строятся, как правило, в виде моделирующего алгоритма, реализации которого приближенного воспроизводит структуру и последовательность протекания анализируемого процесса. Если рассматриваемый процесс подвержен случайным воздействиям, то их влияние имитируется с помощью моделирования случайных объектов с заданными свойствами. При этом искомые характеристики процесса определяются по результатам многократных статистических испытаний. Метод исследования процессов, подверженных слу-

чайным воздействиям, с помощью имитационных моделей принято называть методом статистического моделирования. Основные методы статистического моделирования рассматриваются в подразд. 7.3.

7.2 Метод статистических испытаний (метод Монте-Карло)

Рассмотрим следующую задачу: требуется вычислить определенный интеграл

$$I = \int_0^1 f(x)dx, \tag{7.1}$$

где $f(x)$ - некоторая непрерывная функция, ограниченная на интервале (0,1), $f(0) = 0$, $f(1) = 1$ (рис. 7.1).

Представим себе, что на рис. 7.1 изображено горизонтальное сечение ящика с вертикальными боковыми стенками, внутренней перегородкой и горизонтальным дном.

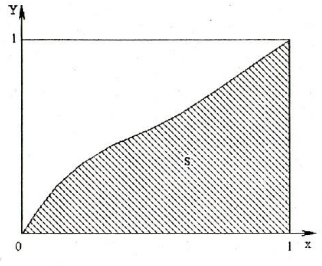


Рис. 7.1

Пусть ящик наполнен песком до высоты h = const. Очевидно, если все песчинки имеют одинаковые размеры, то отношение числа песчинок в отсеке s n_s к общему их числу n_s можно считать приближенно равным отношению площадей s и S, и так как принято S = 1,

$$\frac{n_s}{n_S} = \frac{s}{S} = s \tag{7.2}$$

Но в соответствии с геометрической интерпретацией определенного интеграла его значение эквивалентно площади s. Следовательно

$$I \approx \frac{n_s}{n_S}. \tag{7.2}$$

Предположим далее, что песчинки последовательно бросаются в ящик с помощью устройства, обеспечивающего их равномерное распределение по сечению ящика и будем после каждого бросания определять, попала ли частица в отсек s. Тогда, подставив значения n_s и n_S после заполнения ящика, нетрудно по формуле (7.3) вычислить приблизительную величину I.

Перейдем к построению математической модели рассмотренного процесса. Обозначим координаты i-ой частицы через x_i, y_i (третья координата в нашем случае интереса не представляет). Как отмечалось ранее, для выполнения (7.3) точки (x_i, y_i) должны равномерно располагаться на площади S. Иначе говоря, их координаты могут рассматриваться как реализация независимых случайных величин $\hat{x}$ и $\hat{y}$ равномерно распределенных в интервале (0, 1). Таким образом, если получена последовательность реализаций случайной величины $\hat{\gamma}$, равномерно распределенной в указанном интервале:

$$\gamma_0, \gamma_1, \dots, \gamma_j, \gamma_{j+1}, \dots, \tag{7.4}$$

то акт бросания песчинки можно имитировать, определив координаты точки ее падения равенствами x = γ_j, y = γ_{j+1}. *)

очевидно, что попаданию песчинки в область s соответствует условие x = γ_j, y = γ_{j+1}.

Проведя многократно подобные испытания, можно определить величину отношения $\frac{n_s}{n_S}$ и по формуле (7.2) найти приближенное значение интеграла I.

точности полученных значений интеграла можно воспользоваться рекомендациями по анализу качества оценивания вероятности *p* случайных событий, приведенными в подразд. 4.5. в частности, так как

$$\delta_{\hat{p}^*} = \sqrt{\frac{p(1-p)}{n_s}} \approx \sqrt{\frac{p^*(1-p^*)}{n_S}}, \tag{7.5}$$

где n_S - число испытаний, в соответствии с правилом «трех сигм» искомое значение интеграла «почти наверняка» находится в интервале

$$\left(\frac{n_s}{n_S}\right) - 3\sqrt{\frac{\frac{n_s}{n_S}\left(1-\frac{n_s}{n_S}\right)}{n_S}}, \quad \left(\frac{n_s}{n_S}\right) + 3\sqrt{\frac{\frac{n_s}{n_S}\left(1-\frac{n_s}{n_S}\right)}{n_S}}, \tag{7.6}$$

Пример 1. Вычислить методом Монте-Карло площадь круга единичного радиуса s = $\left(4\int_0^1 dy \int_0^{\sqrt{1-x^2}} dx\right)$.

Как видно на рис. 7.2, площадь круга радиусом R приближенно равна $4\frac{n_s}{n}R_S^2$.

Для определения отношения n_s/n_S при R = 1 с помощью приводимой в подразд. 7.4 формулы (7.16) генерировалась последовательность значений γˆ (расчеты проводились на микрокалькуляторе). В результате ста испытаний (n_S=100) получено n_s = 78. Следовательно, площадь круга единичного радиуса приближенно равна 4 -78/100=3,12 при средней квадратической ошибке 4-√((0,78 0,22)/100) = 0,16. Интересно, что в результате расчетов фактически получено приближенное значение числа π.

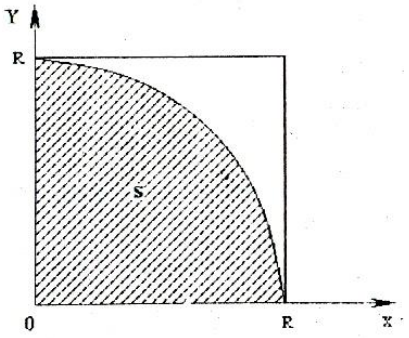


Рис. 7.2

Рассмотренная методика может быть использована и для вычисления определенных интегралов более общего вида:

$$I = \int_a^b f(x)dx,\ 0 \le f(x) \le d. \tag{7.7}$$

В этом случае (рис. 7.3) координатами точек падения песчинок следует считать значения случайных величин, равномерно распределенных в интервалах (a, b) - координата x и $(0, d)$ - координата y. Но, как не трудно показать, случайные величины $\hat{x}$ и $\hat{y}$ связаны с введенной выше случайной величиной $\hat{\gamma}$ простыми соотношениями:

$$\left.\begin{aligned} \hat{x} &= \hat{\gamma}(b-a)+a, \\ \hat{y} &= \hat{\gamma}d, \end{aligned}\right\} \tag{7.8}$$

поэтому, определив по первой из формул (7.8) значение x, соответствующее $\hat{\gamma}_j$, и по второй – значение y, соответствующее γ_{j+1}, можно, как в рассмотренном примере, организовать статистические испытания и найти приближенную величину I:

$$I \approx \frac{n_s}{n_S} S = \frac{n_s}{n_S}(b-a)d. \tag{7.9}$$

Идея осреднения результатов статистических испытаний с использованием специально подобранных случайных величин лежит в основе вычисления методом Монте-Карло различных функционалов и решения систем уравнений (в том числе уравнений в частных производных).

В связи с тем, что быстрота сходимости приближенных значений вычисляемых характеристик к их точным значениям в методе Монте-Карло относительно мала (согласно формулы 7.5 для уменьшения p_(p^*) в 10 раз, необходимо увеличить n_s в сто раз!), удовлетворительное качество моделирования может быть достигнуто, как правило, только при очень большом числе испытаний.

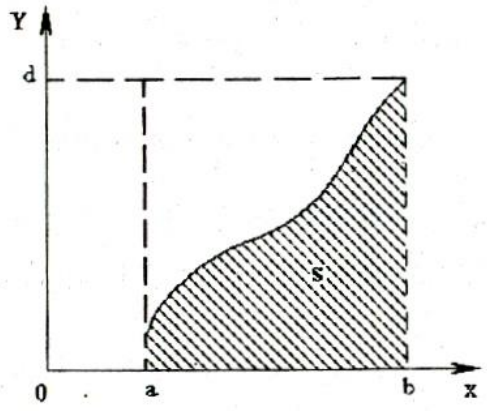


Рис. 7.3

С этой особенностью метода связано то обстоятельство, что, хотя его теоретические основы известны довольно давно, широкое применение метода Монте-Карло как весьма универсального численного метода стало возможным только после появления электронных вычислительных машин.

Разумеется, несмотря на отмеченную универсальность метода Монте-Карло, его практическое применение целесообразно только в ситуациях, когда использование других численных методов менее эффективно (например, при вычислении кратных интегралов).

7.3 Метод статистического моделирования

Метод статистического моделирования предназначен для исследования объектов (процессов, систем), испытывающих случайные воздействия. Поскольку реализация метода основывается на построении последовательностей значений случайных величин и проведении статистических испытаний, за методом статистического моделирования некоторыми авторами сохраняется название метода Монте-Карло. Однако, как будет видно из приводимых ниже примеров, при статистическом моделировании существенна стохастичность анализируемого объекта, в то время как в предыдущем параграфе метод Монте-Карло рассматривался просто как один из методов вычисления функционалов.

Пример 2. Для выполнения программы летной подготовки экипажей самолетов необходимо минимум n суток с метеорологическими условиями П (не обязательно непрерывно следующих друг за другом). Реализацию программы предполагается начать с календарной даты D_0. Для этой и следующих дат известны вероятности осуществления условий П: $p_0, p_1, \dots, p_i, \dots$

Требуется оценить математическое ожидание продолжительности календарного периода, необходимого для выполнения программы.

Прежде всего, отметим, что при $p_0 = p_1 = \cdots p = const$ математическое ожидание продолжительности периода (в сутках) N легко рассчитывается по формуле:

$$\mathrm{M}\left[\widehat{N}\right] = \frac{n}{p}. \tag{7.10}$$

Но, как правило, вероятность условий П имеет хорошо выраженный годовой ход и аналитическое решение задачи становится весьма сложным.

Вот как может быть решена эта задача методом статистического моделирования. Рассматриваем случайную величину $\hat{\gamma}$, равномерно распределенную в интервале (0, 1). По определению, вероятность попадания $\hat{\gamma}$ в интервал (0, p_i) равна p_i. Поэтому при построении имитационной модели реализации программы можно считать, что осуществление условий П в i-е календарные сутки эквивалентно попаданию $\hat{\gamma}$ в указанный интервал (0, p_i).

Пусть получена последовательность значений $\hat{\gamma}$:

$$\gamma_0, \gamma_1, \dots, \gamma_i, \dots, \tag{7.11}$$

и пусть, например, $\gamma_0 > p_0, \gamma_i > p_1, \gamma_2 < p_2, \gamma_3 > p_3,\ \gamma_i < p_i$. Тогда последовательность (7.11) имитирует следующую последовательность погодных условий:

$$\text{П, П, } \overline{\text{П}}\text{, П, …, } \overline{\text{П}}\text{, …,} \tag{7.12}$$

где $\overline{\text{П}}$ означает отсутствие в данные сутки необходимых погодных условий П.

Располагая последовательностью (7.12) достаточной длины, нетрудно подсчитать значение N и, повторив подобные статистические испытания многократно, найти выборочное распределение $\widehat{N}$ и оценку математического ожидания $\widehat{N}$:

$$M^*\left[\widehat{N}\right] = \frac{1}{k}\sum_{i=1}^{k} N_i, \tag{7.13}$$

где k – число испытаний. Анализ качества оценивания $\mathrm{M}\left[\widehat{N}\right]$ выполняется в соответствии с рекомендациями п. 4.6.

заметим, что знание распределения $\widehat{N}$ необходимо для планирования летной подготовки и решения некоторых других оптимизационных задач методом теории статистических решений.

В рассмотренном примере роль случайного фактора играли метеорологические условия. В следующем примере, заимствованном из [2], «природа» случайного фактора может быть любой.

Пример 3. Требуется оценить среднее время безотказной работы прибора, схематически изображенного на рис. 7.4, если известно распределение времени

безотказной работы каждого элемента.

Если бы время безотказной работы каждого элемента t_k было фиксированной величиной, то время безотказной работы прибора t легко было бы рассчитать по формуле

$$t = \min[t_1;\ t_2; \max(t_3;\ t_4): t_s]. \qquad (7.14)$$

(В формуле учтено, что при выходе из строя только элемента 3 или только элемента 4 прибор продолжает работать).

Однако, в действительности время безотказно работы каждого элемента – случайная величина, распределенная по некоторому закону (вообще говоря «своему» для различных элементов), считающемуся известным.

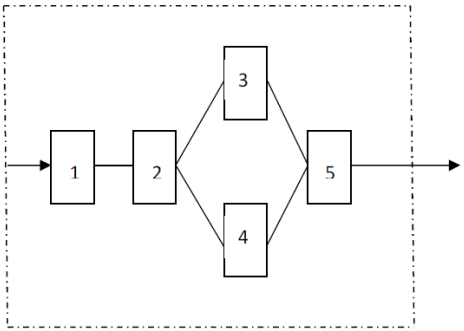


Рис. 7.4

Для построения имитационной модели функционирования прибора рассмотрим последовательности значений случайных величин $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_s$, полученные в соответствии с законами распределения этих величин (вопрос о генерировании таких последовательностей обсуждается в следующем параграфе). Подставив в формулу 7.14 значения первых членов последовательностей, найдем первую реализацию случайной величины $\hat{t}$, затем выполним аналогичные расчеты для вторых членов последовательностей и т. д., после чего, в соответствии с рекомендациями п. 4.6, нетрудно вычислить оценки необходимых параметров распределения $\hat{t}$. В частности,

$$M^*[\hat{t}] = \frac{1}{k}\sum_{i=1}^{k} t_i, \qquad (7.15)$$

где k – число испытаний. *

7.4 Получение последовательностей случайных величин

Методы статистических испытаний и статистического моделирования

предусматривают использование последовательностей значений специально подобранных случайных величин (векторов, событий, процессов, полей). Однако, как будет показано в следующем параграфе, любая из этих последовательностей может быть получена последовательности. В качестве такой последовательности обычно рассматривается последовательность значений случайной величины $\hat{\gamma}$, равномерно распределенной в интервале (0,1).

Для получения реализаций случайной величины $\hat{\gamma}$ (или другой случайной величины, считающейся стандартной) могут быть использованы специальные устройства – генераторы (датчики) случайных чисел.

Простейшими примерами датчиков случайных чисел являются уже упоминавшиеся рулетка или урна с оцифрованными шарами. Процедура получения больших массивов случайных чисел с помощью подобного рода механических датчиков занимают очень много времени и с появлением ЭВМ утратила практическое значение.

В настоящее время для генерирования случайных чисел используются так называемые физические датчики, обеспечивающие получение случайных чисел путем моделирования некоторых физических процессов, имеющих случайный характер (шумы в электронных лампах, радиоактивный распад и т. д.). При генерировании случайных чисел на ЭВМ такие датчики можно рассматривать как специальную приставку к ЭВМ. Полученные значения случайных чисел могут накапливаться в запоминающем устройстве ЭВМ и выдаваться на печать в виде таблицы случайных чисел, которая может в дальнейшем использоваться непосредственно, без обращения к физическому датчику.

Соответствие генерируемых датчиков случайных чисел требуемому закону проверяется с помощью специальных тестов, причем при длительной работе датчика такая проверка должна производится неоднократно. Таблица случайных чисел, естественно, требует только однократной проверки, но обладает ограниченным объемом. С другой стороны, последовательность случайных чисел, генерируемая датчиком, не может быть воспроизведена повторно (что необходимо, например, при откладке программы расчетов).

Таким образом, оба рассмотренных подхода имеют и свои относительные достоинства, и недостатки.

Как показывает сравнительный анализ, результаты которого приведены в табл. 7.1 [1], для решения большинства прикладных задач наиболее приемлем третий метод – метод псевдослучайных числе, к рассмотрению которого и перейдем.

Таблица 7.1

Способ	Достоинства	Недостатки
Таблицы	Проверка однократная. Воспроизводить числа можно	Запас чисел ограничен. Занимает много места в накопителе или медленно вводится. Нужна внешняя память.
Датчики	Запас чисел не ограничен. Сверхбыстрое получение. Места в накопителе не занимает.	Проверка периодическая. Воспроизводить числа нельзя. Требуется специальное внешнее устройство.
Псевдослучайные числа	Проверка однократная. Воспроизводить числа можно. Быстрое получении. Места в накопителе занимает мало. Внешние устройства не нужны.	Запас чисел ограничен.

Псевдослучайными числами называют числа, получаемые по какой-либо формуле (или с помощью какого-либо алгоритмы) и имитирующие значения случайной величины (обычно $\hat{\gamma}$).

Большинство практически используемых для получения псевдослучайных чисел формул имеет вид:

$$\gamma_{j+1} = f(\gamma_j). \tag{7.17}$$

Функция f выбирается таким образом, чтобы точки (γ_j, γ_{j+1}) «достаточно плотно» (при минимальных пропусках) заполняли единичный квадрат. Этому требованию хорошо удовлетворяют функции вида

$$\gamma_{j+1} = \{\mathrm{K}_\gamma\}, \tag{7.18}$$

где К – очень большое число, а $\{z\}$ означает дробную часть числа z. На рис. 7.5 в качестве примера показан график функции (7.18) при К = 21.

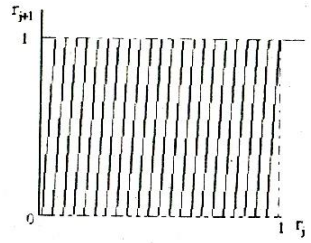


Рис. 7.5

Ниже приводится несколько формул, рекомендуемых для моделирования псевдослучайных чисел на ЭВМ, - формула (7.19) и МК – формулы (7.20) и (7.21)

$$\left.\begin{array}{c}\gamma_j = 2^{-40} m_j, \\ m_{j+1} = 5^{17} m_j (mod\ 2^{40}), \\ m_0 = 0.\end{array}\right\} \tag{7.19}$$

Вторая из формул (7.19) обозначает сравнение по модулю 2^{40} числа $5^{17} m_j$ с числом m_{j+1} (о есть, дает остаток от деления числа $5^{17} m_j$ на 2^{40}), поэтому указанный метод называют методом сравнений или методом вычетов.

$$\begin{array}{l}\gamma_{j+1} = \{37\gamma_j\}, \\ \gamma_{j+1} = \{11\gamma_j + \pi\}.\end{array} \tag{7.20}$$

Последовательности псевдослучайных чисел, получаемые с помощью ЭВМ или МК, вообще говоря, периодические, однако, для рассмотренных формул период настолько велик (при моделировании по формуле (7.21) 8000, а по формуле (7.19) 2^{38}), что это не является помехой при решении большинства прикладных задач. * Пример таблицы псевдослучайных чисел приведен в приложении (табл. 5).

7.5 Преобразования случайных величин

В рассмотренных подразд. 7.2 и 7.3 примерах для проведения статистических испытаний и статистического моделирования использовались последовательности реализаций случайных событий (пример 2), значений случайных величин, имеющих различное распределение (пример 3), и случайных векторов (пример 1). При этом отмечалось, что все такие последовательности могут быть получены путем преобразования последовательности случайных чисел.

Ниже рассматриваются основные методы такого преобразования.

7.5.1 Моделирование последовательностей значений непрерывно распределенных случайных величин

Процесс нахождения значения какой-либо величины $\hat{\xi}$ путем преобразования одного или нескольких значений «стандартной» случайной величины $\hat{\gamma}$ называется разыгрыванием случайной величины $\hat{\xi}$.

Для разыгрывания непрерывно распределенных случайных величин может быть использован метод обратных функций, основанный на следующей теореме.

Если случайная величина $\hat{\xi}$ имеет плотность распределения $\varphi_{\hat{\xi}}(\hat{\xi})$, то распределение случайной величины

$$F_{\hat{\xi}}(\hat{\xi}) = \int_{-\infty}^{\hat{\xi}} \varphi_{\hat{\xi}}(\xi) d\xi \tag{7.22}$$

является равномерным в интервале (0, 1).

Для доказательства учтем, что функция $F_{\hat{\xi}}(\hat{\xi})$ принимает значение $F_{\hat{\xi}}(\xi)$

тогда и только тогда, когда случайная величина $\hat{\xi}$ принимает значение ξ, и поэтому (рис. 7.6)

$$\mathrm{P}\left[F_{\hat{\xi}}(\xi) \le F_{\hat{\xi}}(\hat{\xi}) \le F_{\hat{\xi}}(\xi + \Delta\xi)\right] = P\left[\xi \le \hat{\xi} \le \xi + \Delta\xi\right] \tag{7.23}$$

Но в силу (7.22),

$$P\left[\xi \le \hat{\xi} \le \xi + \Delta\xi\right] = \int_{\xi}^{\xi+\Delta\xi} \varphi_{\hat{\xi}}(\xi)d\xi = F_{\hat{\xi}}(\xi + \Delta\xi) - F_{\hat{\xi}}(\xi), \tag{7.24}$$

и следовательно

$$\mathrm{P}\left[F_{\hat{\xi}}(\xi) \le F_{\hat{\xi}}(\hat{\xi}) \le F_{\hat{\xi}}(\xi + \Delta\xi)\right] = F_{\hat{\xi}}[\xi + \Delta\xi] - F_{\hat{\xi}}(\xi), \tag{7.25}$$

что доказывает справедливость сформулированной теоремы (рис. 7.6).

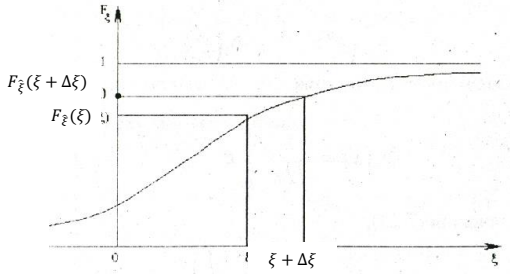


Рис. 7.6

Таким образом, для любой непрерывно распределенной случайной величины $\hat{\xi}$ выполняется:

$$F_{\hat{\xi}}(\hat{\xi}) = \hat{\gamma}. \tag{7.26}$$

На равенстве (7.26) и основан метод обратных функций. Пусть γ_j - значение случайной величины $\hat{\gamma}$, равномерно распределенной в интервале (0, 1). Тогда согласно (7.26), этому значению может быть поставлено в соответствие значение ξ_j разыгрываемой случайной величины $\hat{\xi}$, рассчитанное по формуле:

$$\xi_j = F^{-1}(\gamma_j), \tag{7.27}$$

где F^{-1} - функция, обратная F.

Пример 4. Разыгрывание нормально распределенных случайных величин методом обратных функций.

Начнем с разыгрывания случайной величины $\hat{\xi}^0$, распределенной нормально с параметрами: $M_{\hat{\xi}^0} = 0$; $\delta_{\hat{\xi}^0} = 1$.

Как известно из курса теории вероятностей, функция распределения такой случайной величины имеет вид:

$$F_{\hat{\xi}^0}(\xi^0) = \frac{1}{2} + \Phi_0(\xi^0), \tag{7.28}$$

где

$$\Phi_0(\mathrm{x}) = \frac{1}{\sqrt{2\pi}} \int_0^{\mathrm{x}} \mathrm{e}^{-\frac{t^2}{2}} dt \tag{7.29}$$

Поэтому, на основании (7.27),

$$\xi_j^0 = \Phi_0^{-1}\left(\gamma_j - \frac{1}{2}\right). \tag{7.30}$$

Поскольку значения функции Φ_0^{-1} (или Φ_0) табулированы, формула (7.30) может быть непосредственно использована для разыгрывания случайной величины $\hat{\xi}^0$.

В более общем случае, когда разыгрывается случайная величина $\hat{\xi}$, с математическим ожиданием M и средним квадратическим отклонением δ, достаточно построить последовательность значений $\hat{\xi}^0$, после чего с помощью соотношения перейти к последовательности значений $\hat{\xi}$.

$$\xi_j = \delta\xi_j^0 + \mathrm{M} \tag{7.31}$$

Реализация рассмотренного метода связана с необходимостью определения обратных функций (7.31), что для многих распределений весьма затруднительно. В этих случаях более удобен приводимый метод Неймана.

В основе метода Неймана лежат следующие соображения. Разыгрывается случайная величина $\hat{\xi}$, определенная на конечном интервале (a, b). Известна плотность распределения $\hat{\xi}$: $\varphi_{\hat{\xi}}(\xi)$, ограниченная некоторым числом c.

Но определению, вероятность попадания значений $\hat{\xi}$ в интервал (a_1, b_1)

$$\mathrm{P}(a_1 \le \hat{\xi} \le b_1) = \int_{b_1}^{a_1} \varphi_{\hat{\xi}}(\xi) d\xi \tag{7.32}$$

соответствует площади s на рис. 7.7. Руководствуясь этим условием, и подбирают моделирующую случайную величину.

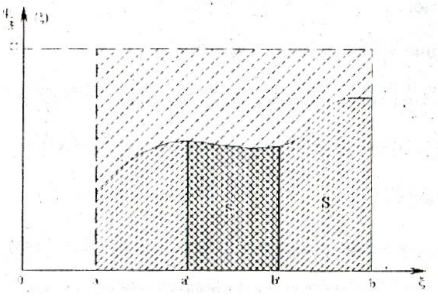


Рис. 7.7

Рассмотрим множество случайных точек с координатами $\hat{\gamma}'$ и $\hat{\gamma}''$, равномерно распределенными соответственно в интервалах (a, b) и (0, c), и выделим из него подмножество точек, расположенных ниже кривой $\varphi_{\hat{\xi}}(\xi)$. Обозначим число таких точек через n_s, а число точек, попавших в область s через n_s. Очевидно, что отношение $\frac{n_s}{n_S}$ с увеличением числа точек стремиться к отношению $\frac{s}{S}$, то есть, поскольку S = 1, $\frac{n_s}{n_S} \to s$.

Так как, далее, координата $\hat{\gamma}''$ играет только роль индикаторов при отборе «нужного» подмножества точек, из сказанного вытекает простой алгоритм разыгрывается случайной величины $\hat{\xi}$:

генерируется первое случайное число γ_0;

по формуле

$$\hat{\gamma}' = a + \hat{\gamma}(b + a) \qquad (7.32)$$

находится значение γ_0' случайной величины $\hat{\gamma}'$,

вычисляется $\varphi(\gamma_0')$;

генерируется второе случайной число γ_1;

по формуле: $\hat{\gamma}'' = \hat{\gamma}$c

находится значение γ_1'' случайной величины $\hat{\gamma}_1''$;

сравниваются значения $\varphi(\gamma_0')$ и $\hat{\gamma}_1''$ если $\varphi(\gamma_0') < \hat{\gamma}_1''$, переходят к рассмотрению следующей пары случайных чисел, если $\varphi(\gamma_0') \geq \hat{\gamma}_1''$, принимается, что γ_0' имитирует ξ_0;

строится последовательность отобранных таким образом значений $\hat{\gamma}_1'$, которая и является моделью последовательности значений ξ_0.

7.5.2 Моделирование последовательностей значений дискретных случайных величин

Пусть требуется разыграть случайную величину $\hat{\xi}'$, распределение которой представлено в табл. 7.2.

Таблица 7.2

Значения $\hat{\xi}$	ξ_1	ξ_2	…	ξ_n
Вероятности	p_1	p_2	…	p_n

Рассмотрим последовательность случайных чисел:

$$\gamma_0, \gamma_1, \dots, \gamma_j, \dots$$

Так как случайная величина $\hat{\gamma}$ распределена равномерно в интервале (0, 1), вероятность попадания ее значений в интервал (0, p_1), равна p_1, в интервал (p_1, $p_1 + p_2$) - p_2 и т. д. (рис. 7.8).

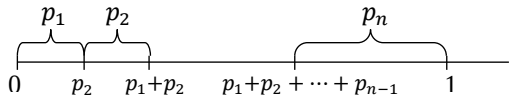


Рис. 7.8

Поэтому осуществление значения $\hat{\gamma}_0$, лежащего в интервале $(0, p_1)$, имитирует осуществление значения $\hat{\xi}$, равного ξ_1, значения $\hat{\gamma}$, лежащего в интервале $(p_1, p_1 + p_2)$ – осуществление ξ_2 и т. д.

Пример 5. Разыграть случайную величину $\widehat{N}'$ - общее количество облачности (в баллах), заданную табл. 7.3.

Таблица 7.3

N_i	0	1	2	3	4	5	6	7	8	9	10
p_i	0,27	0,12	0,05	0,11	0,02	0,08	0,04	0,01	0,06	0,06	0,18

В табл. 7.4 приводятся первые десять случайных чисел, взятых из таблицы случайных чисел табл. 5 (Приложение), и определяются соответствующие значения $\widehat{N}$.

Таблица 7.4

γ_i	0,10097	0,32533	0,76520	0,13586	0,34673	0,54876
N_i	0	1	9	0	1	3
γ_i	0,80959	0,09117	0,39292	,74945		
N_i	9	10	1	8		

Таким образом смоделирована следующая последовательность значений $\widehat{N}$:0, 1, 9, 0, 1, 3, 9, 10, 1, 8.

7.5.3 Моделирование последовательностей случайных событий

Задача моделирования последовательностей случайных событий легко сводится к разыгрыванию соответствующих дискретных случайных величин. Действительно, пусть $\widehat{\mathrm{A}}_1$, $\widehat{\mathrm{A}}_2$, …, $\widehat{\mathrm{A}}_n$ – попарно независимые случайные события, образующие полную группу. Известны вероятности осуществления этих событий: $p_1, p_2, \ldots, p_n (\sum_{i=1}^{n} p_i = 1)$.

Закодируем факт осуществления события $\widehat{\mathrm{A}}_1$ единицей, события $\widehat{\mathrm{A}}_2$ - двойкой и т. д. и разыграем дискретную случайную величину $\hat{\xi}$, заданную табл. 7.5.

Таблица 7.5

Значения $\hat{\xi}$	1	2	…	n
Вероятности	p_1	p_2	…	p_n

В случайной последовательности значений $\hat{\xi}$ остается произвести замену $\hat{\xi}_i$ на A_i.

7.5.4 Моделирование последовательностей значений случайных векторов

Моделирование последовательностей значений вектора $\widehat{X}_{<n>}$, компонентами которого являются независимые случайные величины $\hat{x}_1$, $\hat{x}_2$, …, $\hat{x}_n$ - сводится к разыгрыванию этих случайных величин в соответствии с законами их распределения. Процедура моделирования может быть организована следующим образом.

Генерируется последовательность случайных чисел:

$$\gamma_1, \gamma_2, \dots, \gamma_n, \tag{7.34}$$

Рассмотрим первые n членов последовательности (7.34). пользуясь методами разыгрывания случайных величин, приведенными в подразд. 7.5.1 и 7.5.2, рассчитаем значения случайных величин $\hat{x}_1$, $\hat{x}_2$, …, $\hat{x}_n$, определяющие первое значение вектора $\widehat{X}_{<n>}$, затем по следующему «отрезку2 последовательности (7.34) длиной n найдем второе значение $\widehat{X}_{<n>}$ и т. д.

Примеры моделирования двумерного случайного вектора приводились в подразд. 7.2 (пример 7.1) и подразд. 7.3 (пример 3).

В общем случае, когда между компонентами вектора $\widehat{X}_{<n>}$ существует статистическая связь, процедура моделирования несколько усложняется. Вначале выбирается и разыгрывается какая-либо из случайных величин – компонент вектора $\widehat{X}_{<n>}$, например, $\hat{x}_1$. Затем для первого полученного значения этой величины находится условное распределение второй случайной величины, например $\hat{x}_2$, и в соответствии с этим распределением моделируется ее значение. Аналогично разыгрывается третья случайная величина при полученных ранее первых двух случайных величин и т. д.

Пример 6. В табл. 7.6 приведено распределение случаев одновременного существования на двух близко расположенных друг к другу станциях «малооблачно» (N) и «облачной» ($\overline{N}$) погоды. Требуется построить последовательность значений вектора $\langle \widehat{N}_1, \widehat{N}_2 \rangle$.

Таблица 7.6

Значения $\widehat{N}_2$	Значения $\widehat{N}_1$		Σ
	N	$\overline{N}$	
N	250	200	450
$\overline{N}$	150	400	550
Σ	400	600	1000

Оценим вероятности N и $\overline{N}$ на станции 1:

$$p^*(\widehat{N}_1 = N) = \frac{400}{1000} = 0{,}400,$$
$$p^*(\widehat{N}_1 = \overline{N}) = \frac{600}{1000} = 0{,}600, \qquad (7.35)$$

И условные вероятности осуществления этих условий на станции 2.

$$p^*(\widehat{N}_2 = N/\widehat{N}_1 = N) = \frac{250}{400} = 0{,}625;$$
$$p^*(\widehat{N}_2 = \overline{N}/\widehat{N}_1 = N) = \frac{150}{400} = 0{,}375;$$
$$p^*(\widehat{N}_2 = N/\widehat{N}_1 = \overline{N}) = \frac{200}{600} = 0{,}333; \qquad (7.36)$$
$$p^*(\widehat{N}_2 = \overline{N}/\widehat{N}_1 = \overline{N}) = \frac{400}{600} = 0{,}667.$$

Пусть генерирована та же последовательность случайных чисел, что и в примере 5 (подразд. 7.5.2):

γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
0,70097;	0,32533;	0,76520;	0,13586;	0,34673;	0,54876;
γ_7	γ_8	γ_9	γ_{10}		
0,80959;	0,09117;	0,39292;	0,74945;		

Начнем со случайной величины $\widehat{N}_1$. В соответствии с (7.35) будем считать, что $\widehat{N}_1 = N$, если $\hat{\gamma} = \gamma \leq 0{,}400$. Тогда γ_1 = 0,10097 имитирует значение $\widehat{N}_1 = N$. Теперь рассмотрим второе значение γ. Согласно первой из формул (7.37) следует положить $\widehat{N}_2 = N$ (поскольку $\gamma_2 < 0{,}625$). Таким образом, первая реализация случайного вектора равна $< N, N >$. *) Аналогично получим, что пара случайных чисел γ_3, γ_4 имитирует реализацию случайного вектора $\langle \overline{N}, N \rangle$ и т. д.

8 ОТБОР ПРЕДИКТОРОВ ПРИ ПОСТРОЕНИИ ПРОГНОСТИЧЕСКИХ МОДЕЛЕЙ

8.1 Постановка задачи

Разработка статистических методов метеорологических прогнозов начинается с определения подлежащих прогнозу характеристик состояния атмосферы – предиктантов и характеристик, предположительно влияющих на значения предиктантов и известных к моменту составления прогноза – предикторов.

Назначение предиктантов производится в соответствии с требованиями потребителей прогностической информации, в роли которых могут выступать как специалисты соответствующих подразделений народнохозяйственных или военно-технических систем, так и сами метеорологи. В последнем случае предсказанные значения предиктантов используются в дальнейшем в качестве вторичных предикторов для составления прогнозов в интересах внешних потребителей. Так обстоит дело, например, при составлении авиационных прогнозов погоды на основании результатов предварительного прогноза синоптического положения.

Формирование перечня предполагаемых (виртуальных) предикторов выполняется метеорологом с учетом закономерностей атмосферных процессов, определяющих, по его мнению, значения предикторов. При этом, стремясь наиболее полно отразить в модели указанные закономерности, метеоролог часто включает в предварительный перечень и предикторы, не имеющие очевидной тесной связи с предиктантом, рассчитывая на то, что такое включение может привести только к повышению или, в крайнем случае, к сохранению модели.

Однако при построении статистических моделей на материале архивных выборок ограниченного объема утверждение о допустимости произвольного расширения перечня используемых предикторов справедливо лишь для обучающих частей выборок. Включение в модели дополнительных предикторов, действительно, не может привести к снижению качества «прогнозов» по этим частям выборки. Но существенно по-другому обстоит дело с прогностической эффективностью моделей, характеризуемой качеством прогнозов для случаев, не вошедших в обучающую выборку. В отличие от «прогнозов» по обучающей выборке качество прогнозов по экзаменационным выборкам не являются монотонной функцией числа включаемых в модель предикторов из предварительного перечня.

На рис. 8.1 и 8.2 приводятся средние квадратические ошибки линейной регрессии, построенной для прогноза максимальной температуры воздуха по данным табл. 5.2. В предварительный перечень было включено 8 предикторов из

которых рассматриваемым в подразд. 8.3 методом последовательного включения образовывались лучшие группы из k предикторов (k = 1, …8). Очередные предикторы, добавляются в ранее сформированные группы, указаны на горизонтальных осях графиков. В целях большей наглядности точки на графиках соединены отрезками.

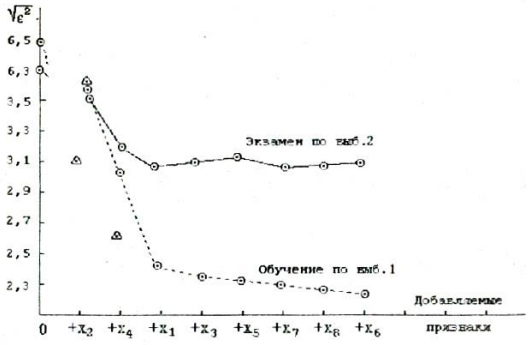


Рис. 8.1

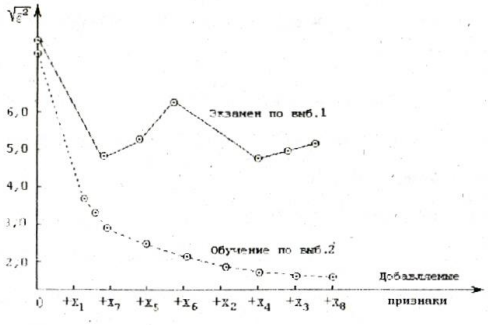


Рис. 8.2

Как видно на рис. 8.1, где представлены результаты испытаний моделей, построенных по выборке 1 (m = 40), добавление в группу из трех предикторов (x_2, x_4 и x_1) лучшего из оставшихся предикторов (x_3) приводит не к уменьшению, а к увеличению средних квадратических ошибок экзаменационных прогнозов. Еще более отчетливо указанный эффект наблюдается для моделей, построенных по выборке 2 (m = 20) (рис. 8.2). оптимальная группа в этом случае состоит всего из одного предиктора x_1.

Рассмотренные примеры иллюстрируют типичную в прогностической практике ситуацию, когда при формировании оптимальной группы предикторов необходим их отбор из предварительного перечня. Для проведения такого отбора должно быть установлены правила сравнительного анализа качества моделей с

различными сочетаниями предикторов (критерии отбора) и процедуры генерирования этих сочетаний.

Начнем с рассмотрения наиболее употребительных критериев отбора.

8.2 Критерии отбора предикторов

Показатели качества сравниваемых моделей должны характеризовать их прогностическую эффективность, определяемую для генеральной совокупности. В метеорологической практике к таким показателям принято относить:

для количественных прогнозов: математическое ожидание квадрата или абсолютной величины ошибки прогнозов, множественный коэффициент корреляции (коэффициент корреляции между предсказанными и осуществившимися значениями предиктанта) и др.;

для альтернативных прогнозов: вероятность ошибочных прогнозов, показатели Н. А. Багрова, А. М. Обухова, количество прогностической информации и др.

При разработке методов специализированных прогнозов в интересах конкретных потребителей качество моделей должно характеризоваться показателями эффективности функционирования этих потребителей (при условии оптимального использования ими прогностической информации).

Лучшей из сравниваемых моделей (сочетаний предикторов) признается модель, доставляющая соответствующий экстремум выбранному показателю.

В связи с тем, что построение моделей производится на материале ограниченных выборок, метеоролог при сравнении качества моделей вынужден оперировать с оценками показателей качества. Поэтому при сравнении по показателю К двух моделей первой и второй вопрос о предпочтительности одной из них, например модели 1, может быть решен только в вероятностном смысле. По сути говоря, задача сводится к статистической проверке гипотезы H_0: $К_1 > К_2$ (или $К_1 < К_2$).

Для решения указанной задачи, очевидно, необходимо знание закона распределения выборочных оценок показателя К. поскольку разработчику моделей этот закон, как правило, неизвестен, то отбор предикторов приходится основывать на некоторых правдоподобных предположениях.

Прежде всего, при сравнении сочетаний с одинаковым количеством предикторов предполагается, что лучшему сочетанию соответствует экстремальное значение выборочной оценки рассматриваемого показателя качества. Для выборок, объем которых во много раз превосходит число сравниваемых сочетаний, это предположение представляется достаточно обоснованным. Но при относительно не большом объеме выборок выводы о сравнительной прогностической

эффективности моделей, основанные на анализе оценок показателя качества, часто оказываются ошибочными.

Так, если при выборе лучшей из однокомпонентных регрессионных моделей, построенных по данным табл. 6.2 может основываться на результатах сравнения оценок $r^{*2}_{\hat{x}_0\hat{x}_1}$ (i=1, 2, …, 8), полученных для первой части выборки (m=40), предпочтение должно быть отдано регрессии $\hat{x}_1$ на $\hat{x}_2$, в то время как при оценивании $r^{*2}_{\hat{x}_0\hat{x}_1}$ по полной выборке (m=60) лучшей следует признать регрессию $\hat{x}_0$ на $\hat{x}_1$.

Несмотря на отмеченный недостаток, предположение о допустимости замены показателей качества моделей соответствующими выборочными оценками широко используется в прикладной статистике при сравнении моделей с одинаковым числом предикторов. В то же время, это предположение становится совершенно неприемлемым при сравнении прогностической эффективности моделей с различным количеством предикторов.

Действительно, если, например, качество моделей характеризуется значением множественного коэффициента корреляции R, то, поскольку R = 0 математическое ожидание квадрата выборочных оценок R* подчинено закону [6]:

$$\mathrm{M}\left[\hat{R}^{*2}\right] = \frac{n}{m-1}, \tag{8.1}$$

при фиксированном объеме выборки m максимальные значения R^{*2} могут быть получены для моделей со сравнительно низкой прогностической эффективностью при включении в них достаточно большого числа предикторов.

Вероятность подобного рода ошибок несколько уменьшается, если при сравнении регрессионных моделей вместо R^{*2} использовать статистику

$$R^{*2}_k = \left(1 - \frac{m-1}{m-n}\right)(1 - R^{*2}), \tag{8.2}$$

причем лучшим считать сочетание предикторов, доставившее максимум R^{*2}_k*

однако и выводы, основанные на сравнении значений R^{*2}_k, для моделей с различным количеством предикторов нередко оказываются неверными. Например, если для сравнения одно- и восьмикомпонентных моделей воспользоваться данными, представленными на рис. 8.2 (обучающая выборка), то по оценкам не только R^{*2}, но и R^{*2}_k, предпочтение должно быть отдано восьмикомпонентной выборке.

В общем, к аналогичным выводам приводит и практика использования всех других показателей эффективности прогностических моделей: отбор непосредственно по выборочным оценкам этих показателей имеет смысл только среди моделей с одинаковым количеством предикторов.

При сравнении моделей с различным количеством предикторов необходим анализ их прогностической эффективности на независимом статистическом материале. В этих целях, помимо основной обучающей отбираются лучшие из моделей с одинаковым количеством предикторов, формируется дополнительная экзаменационная (контрольная) выборка, по которой оцениваются показатели качества моделей, отобранных по обучающей выборке, и окончательно определяется оптимальное сочетание предикторов.

Деление архивной выборки на обучающую и экзаменационную части может производиться двумя способами. Первый из них состоит в образовании на архивном материале двух непересекающихся выборок постоянного состава. При этом отношение объемов обучающей и экзаменационной выборок обычно принимается близким к 2:1. Второй способ предусматривает многократный обмен вариантами между обучающей и экзаменационной выборками (метод скользящего экзамена). В основном варианте метода скользящего экзамена в экзаменационную часть первоначально включаются только значения предиктанта и предикторов, полученных при наблюдении 1. Поэтому одному варианту оцениваются ошибки прогнозов, составленные с использованием различных моделей, построенных по остальным вариантам. Затем вариант 1 возвращается в обучающую часть выборки, а вариант 2 передается из этой части в экзаменационную, вновь строятся модели и оценивается качество прогнозов и т. д. После окончания перебора для каждого класса моделей с фиксированным числом предикторов рассчитывается среднее (m испытаниям) значение принятого показателя качества, например $\overline{\delta^2_{\text{экз}}}$, и определяется оптимальное число предикторов $k_{\text{опт}}$ числом предикторов, лучшая из которых и считается окончательной.

8.3 Процедуры отбора предикторов

Рассмотрим следующую простую ситуацию. Пусть из n предикторов предварительного перечня требуется отобрать один лучший, считая критерием отбора максимум квадрата коэффициента корреляции $r^2_{\hat{x}_0\hat{x}_1}$ между предиктантом x_0 и предиктором x_i (i=1, 2, …, n).

Как отмечалось выше, лучшим в этой ситуации должен быть признан предиктор с максимальным значением $r^2_{\hat{x}_0\hat{x}_1}$. Но выборочные оценки $r^2_{\hat{x}_0\hat{x}_1}$ практически никогда не равны «истинным» величинам $r^2_{\hat{x}_0\hat{x}_1}$. При этом, поскольку выбирается предиктор с максимальной оценкой, более вероятно, что эта оценка завышена ($r^{*2} > r^2$), а не занижена.

В силу сказанного, реализация рассмотренного алгоритма может привести, во-первых, к искусственному «вздутию» коэффициента корреляции, в результате чего прогностическая эффективность отобранной модели окажется ниже

ожидаемой, и, во-вторых, отобранный предиктор «в действительности» (по величине r^2) не является лучшим.

Очевидно, что вероятность отмеченных ошибок тем больше, чем больше предикторов включено в предварительный перечень. Поэтому чрезмерное расширение предварительного перечня в расчете на то, что «лишние» предикторы все равно будут исключены при построении моделей, недопустимо. *

Пусть теперь в той же исходной ситуации требуется отобрать лучшую из всех (а не только однокомпонентных) моделей. Общее число таких моделей (т. е. число возможных сочетаний из n элементов), включая модель, не содержащую предикторов, равно 2^n. Следовательно, если в предварительный перечень включено 10 предикторов (случай, не исключительный в метеорологической практике), число возможных моделей равно $2^{10} = 1024$. Таким образом, уже при нескольких предполагаемых предикторах число подлежащих сравнению выборочных оценок показателя качества для всех возможных моделей настолько велико (много больше n), что указанные выше ошибки отбора становятся практически неизбежным.

Выход из этого положения состоит в использовании специальных процедур отбора, предусматривающих сокращение количества сравниваемых сочетаний предикторов за счет исключения из рассмотрения моделей, которые априори признаются неоптимальными.

Наибольшее распространение в прогностической практике получили пошаговые процедуры формирования анализируемых сочетаний предикторов: процедура последовательного присоединения предикторов и процедура последовательного исключения предикторов.

Процедура последовательного присоединения предикторов используется в тех случаях, когда, исходя из опыта построения аналогичных моделей, следует ожидать, что оптимальной окажется модель, содержащая всего несколько предикторов из предварительного перечня.

Процедура предусматривает следующую последовательность операций.

На первом шаге по обучающей выборке строятся модели, содержащие по одному предиктору, и по результатам сравнения оценок показателя качества определяется лучшая модель.

На втором шаге анализируются модели, содержащие по два предиктора, один из которых, отобранный на первом шаге. Вновь по обучающей выборке определяется лучшая из рассматриваемых (двух компонентных) моделей.

На следующих шагах последовательно рассматриваются трех-, четырехкомпонентные модели и т. д., пока не будут использованы все предикторы или не будет принято решение об окончании отбора. Такое решение принимается по

результатам прогнозов, составляемых с помощью «лучших» моделей на каждом шаге отбора для экзаменационной выборки. Отбор заканчивается, когда на следующем шаге качество прогнозов, судя по оценкам показателя качества $К^{*}_{экз}$, перестает повышаться.

Таким образом, на первом шаге отбора рассматривается n сочетаний предикторов, на втором – (n – 1) и т. д. Максимальное количество рассмотренных сочетаний при реализации процедур последовательного присоединения равно n + (n – 1) + … + 1 = $\frac{n+1}{2}n$, то есть исключенными из рассмотрения (при n> 1) оказываются, по крайней мере, ($2^{n}\frac{n+1}{2}n$) сочетаний. Если. Например, n = 8, то не рассматривается не менее ($2^{8}\frac{8+1}{2}8$) = 988 сочетаний.

Процедура последовательного исключения предикторов рекомендуется в тех случаях, когда, исходя из опыта построения аналогичных моделей, следует ожидать, что оптимальной окажется модель, содержащая большинство предикторов предварительного перечня.

На первом шаге по обучающей выборке строится модель, включающая все n предполагаемых предикторов. Прогностическая эффективность модели характеризуется оценкой показателя качества, рассчитываемого по экзаменационной выборке.

На втором шаге из рассмотрения поочередно исключаются первый, второй и т. д. предикторы, на материале обучающей выборки строятся n моделей, содержащих по (n – 1)-му предиктору, и по величине $К_{об}$ отбирается лучшая модель.

На третьем шаге из состава предикторов лучшей модели, определенной на втором этапе, вновь поочередно исключается по одному предиктору, строятся (n – 1) моделей, содержащих по (n – 2) предиктора, и т. д.

После определения лучшей модели на каждом шаге отбора с ее помощью составляются прогнозы для экзаменационной выборки и рассчитываются значения $К^{*}_{экз}$, следующий шаг не приводит к повышению качества моделей.

Разумеется, ни процедура последовательного присоединения, ни процедура последовательного исключения предикторов (впрочем, как и другие используемые в прикладной статистике процедуры) не гарантируют безошибочный выбор оптимальной модели. Однако, как показывает прогностическая практика, их использование приводит к существенно лучшим результатам, чем полный перебор всех сочетаний предполагаемых предикторов.

Пример 3. Пользуясь данными табл. 6.2, построить уравнение линейной среднеквадратической регрессии для прогноза максимальной температуры воздуха.

В соответствии с постановкой задачи показателем качества модели следует

считать математическое ожидание квадрата ошибки регрессии. Поскольку объем имеющегося архивного материала (m = 60) относительно невелик (n = 8!), оптимальная модель будет включать, вероятно, лишь небольшую часть предполагаемых предикторов и целесообразно воспользоваться процедурой последовательного присоединения предикторов.

Рассмотрим два варианта реализации этой процедуры: при фиксированном составе обучающей и экзаменационной выборок и при обменен вариантами между ними (метод скользящего экзамена).

Первый вариант решения. Выполним деление выборки на обучающую и экзаменационные части так, как это показано в табл. 6.2.

При выборке лучшей из однокомпонентных моделей можно воспользоваться данными табл. 6.4, в соответствии с которыми максимальное значение $\bar{\varepsilon}^2_{\text{об}}$ достигается для модели с предиктором x_2. Соответствующее уравнение регрессии имеет вид:

$$\hat{\hat{x}}_0 = 10{,}082 + 1{,}116\hat{x}_2. * \qquad (8.3)$$

Составив по формуле (8.3) прогнозы для экзаменационной выборки, найдем:

$$\sqrt{\overline{\varepsilon^2_{\text{экз}}}} = 3{,}475.$$

Для выбора «лучшей пары» к предиктору x_2 можно по формулам типа (6.36) рассчитать значения частных коэффициентов корреляции $r_{\hat{x}_0,\hat{x}_1,\hat{x}_2}$ (I = 1,3,4,…,8), судя по которым к предиктору x_2 целесообразно присоединить предиктор x_4, и лучшей является модель

$$\hat{\hat{x}}_0 = 12{,}394 + 1{,}074\hat{x}_2 - 0{,}553\hat{x}_4. \qquad (8.4)$$

Средняя квадратическая ошибка прогнозов по формуле (8.4) для вариантов экзаменационной выборки равна

$$\sqrt{\overline{\varepsilon^2_{\text{экз}}}} = 3{,}152.$$

Поскольку эта ошибка меньше, чем для однокомпонентной регрессии, продолжим отбор.

Построим уравнение регрессии для всех троек предикторов, включающих x_2 и x_4, и найдем $\bar{\varepsilon}^2_{\text{об}}$. В соответствии с результатами расчетов лучшей является регрессии

$$\hat{\hat{x}}_0 = 10{,}964 + 0{,}487\hat{x}_2 - 0{,}615\hat{x}_4 + 0{,}639\hat{x}_1. * \qquad (8.5)$$

Составив по выражению (8.5) прогнозы для экзаменационной выборки, найдем:

$$\sqrt{\overline{\varepsilon^2_{\text{экз}}}} = 3{,}044.$$

Переходим к рассмотрению сочетаний из четырех предикторов, содержащих x_2, x_4 и x_1. Построим уравнение регрессии для каждого сочетания и рассчитав средние квадратические ошибки по обучающей выборке, определим лучшее уравнение:

$$\hat{\hat{x}}_0 = 9{,}295 + 0{,}470\hat{x}_2 - 0{,}446\hat{x}_4 + 0{,}645\hat{x}_1 + 0{,}232\hat{x}_3. \quad (8.6)$$

*

которому соответствует средняя квадратическая ошибка прогнозов по экзаменационной выборке:

$$\sqrt{\overline{\varepsilon^2_{\text{экз}}}} = 3{,}064.$$

Так кА на этом шаге произошел рост $\overline{\varepsilon^2_{\text{экз}}}$, для оперативного использования следует рекомендовать уравнение (8.5).

Второй вариант решения. Для выбора лучшей однокомпонентной модели включим вначале в обучающую выборку наблюдения с номерами 2, 3, …, 60, построим уравнения регрессии и найдем ошибки прогноза по каждому уравнению для наблюдения 1: $\varepsilon_1(1)$, $\varepsilon_2(1)$, …, $\varepsilon_8(1)$.

Затем вернем в обучающую выборку наблюдение 1, выделив для экзамена наблюдение 2, снова построим уравнения регрессии и вычислим $\varepsilon_1(2)$, $\varepsilon_2(2)$, …, $\varepsilon_8(2)$.

Проделав аналогичные операции для остальных случаев перераспределения наблюдений, найдем суммы $\sum_{i=1}^{60} \varepsilon^2(j)$ для моделей, включавших предиктор $x_1(j = 1)$, предиктор $x_2(j = 2)$ и т. д. Как показывают результаты расчетов, лучшими по значению этого показателя являются регрессии $\hat{x}_0$ на $\hat{x}_1\left(\sqrt{\frac{1}{60}\sum_{i-1}^{60} \varepsilon_j^2\,(1)} = 3{,}529\right)$.

Теперь применим метод скользящего экзамена для определения лучшей из пар предикторов, включающих x_1. Использовав тот же алгоритм, что и для отбора лучшего предиктора, найдем оптимальное сочетание предикторов x_1, x_4, для которого соответствующая средняя квадратическая ошибка равна 3,010.

Аналогично найдем лучшее сочетание из трех предикторов, включающее x_1 и x_4: x_1, x_4, x_2, со средней квадратической ошибкой 2,699.

Для лучшего сочетания из четырех предикторов (x_1, x_4, x_2 и x_7) соответствующая ошибка равна 2,681, а для лучшего сочетания из пяти предикторов (x_1, x_4, x_2, x_7 и x_6) она равна 2,708. Таким образом, лучшим должно быть признано сочетание из четырех предикторов: x_1, x_4, x_2 и x_7. Уравнение регрессии с этими предикторами, полученное по полной выборке, имеет вид:

$$\hat{\hat{x}}_0 = 12{,}141 + 0{,}530\hat{x}_1 + 0{,}518\hat{x}_2 - 0{,}500\hat{x}_4 - 0{,}165\hat{x}_7. \quad (8.7)$$

Это уравнение и должно быть рекомендовано для составления оперативных прогнозов. Как видно из сравнений формул (8.7) и (8.5), более полное использование информации, содержащейся в архивной выборке, при скользящем экзамене позволило увеличить оптимальное число предикторов и соответственно несколько повысило качество модели.

ПРИЛОЖЕНИЕ

Таблица 1

Значение функции $\phi_{0(x)} = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2az}}$

0,0	0,00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0,6	22575	22907	23237	23565	23891	24251	24537	24857	25175	25490
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0,8	28814	20103	29389	29673	29955	30234	30511	30785	31057	31327
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214
1,1	36443	36650	36864	37076	37286	37493	37698	37900	38100	38298
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3,0	0,49865		3,1	49903	3,2	49931	3,3	49952	3,4	49966
3,5	49977		356	49984	3,7	49989	3,8	49993	3,9	49995
4,4	499968									
5,0	49999997									

Таблица 4

Значение функции $f_{\theta;k_1;k_2} = F^{-1}_{\hat{f}(k_1;k_2)}(1-\theta)$

k_2		k_1 – степеней свободы числителя											
		1	2	3	4	5	6	7	8	9	10	11	12
k_2 - степени свободы знаменателя	1	161 4052	200 4999	216 5403	225 5625	230 5764	234 5889	237 5 928	239 5 981	241 6022	242 6056	243 6082	244 6106
	2	18,51 98,49	19,00 99,01	19,16 99,17	19,25 99,25	19,30 99,30	19,33 99,33	19,36 99,34	19,37 99,36	19,38 99,38	19,39 99,40	19,40 99,41	19,41 99,42
	3	10,13 34,12	9,55 30,81	9,28 29,46	9,12 28,71	9,01 28,24	8,94 27,91	8,88 26,67	8,84 27,49	8,81 27,34	8,78 27,23	8,76 27,13	8,74 27,05
	4	7,71 21,20	6,94 18,00	6,59 16,69	6,39 15,98	6,26 15,52	6,16 15,21	6,09 14,98	6,04 14,80	6,00 14,66	5,96 14,54	5,93 14,45	5,91 14,37
	5	6,62 16,26	5,79 13,27	5,41 12,06	5,19 11,39	5,05 10,97	4,95 10,67	4,88 10,45	4,82 10,27	4,78 10,15	4,74 10,05	4,70 9,96	4,68 9,89
	6	5,99 13,74	5,14 10,92	4,76 9,78	4,53 9,15	4,39 8,75	4,28 8,47	4,21 8,26	4,15 8,10	4,10 7,98	4,06 7,87	4,03 7,79	4,00 7,72
	7	5,59 12,25	4,74 9,55	4,35 8,45	4,12 7,85	3,97 7,46	3,87 7,19	3,79 7,00	3,73 6,84	3,68 6,71	3,63 6,62	3,60 6,54	3,57 6,47
	8	5,32 11,26	4,46 8,65	4,07 7,59	3,84 7,01	3,69 6,63	3,58 6,37	3,50 6,19	3,44 6,03	3,39 5,91	3,34 5,82	3,31 5,74	3,28 5,67
	9	5,12 10,56	4,26 8,02	3,86 6,99	3,63 6,42	3,48 6,06	3,37 5,80	3,29 5,62	3,23 5,47	3,18 5,35	3,13 5,26	3,10 5,18	3,07 5,11
	10	4,96 10,04	4,10 7,56	3,71 6,55	3,48 5,99	3,33 5,64	3,22 5,39	3,14 5,21	3,07 5,06	3,02 4,95	2,97 4,85	2,94 4,78	2,91 4,71
	11	4,84 9,85	3,98 7,20	3,59 6,22	3,36 5,67	3,20 5,32	3,09 5,07	3,01 4,88	2,95 4,74	2,90 4,63	2,86 4,54	2,82 4,46	2,79 4,40
	12	4,74 9,33	3,88 6,93	3,49 5,95	3,26 5,41	3,11 5,06	3,00 4,82	2,92 4,65	2,85 4,50	2,80 4,39	2,76 4,30	2,72 4,22	2,69 4,16
	13	4,67 9,07	3,80 6,70	3,41 5,74	3,18 5,20	3,02 4,86	2,92 4,62	2,84 4,44	2,77 4,30	2,72 4,19	2,67 4,10	2,63 4,02	2,60 3,96
	14	4,60 8,86	3,74 6,51	3,34 5,56	3,11 5,03	2,96 4,69	2,85 4,46	2,77 4,28	2,70 4,14	2,65 4,03	2,60 3,94	2,56 3,86	2,53 3,80
	15	4,54 8,68	3,68 6,36	3,29 5,42	3,06 4,89	2,90 4,56	2,79 4,32	2,70 4,14	2,64 4,00	2,59 3,89	2,55 3,80	2,51 3,73	2,48 3,67
	16	4,49 8,53	3,63 6,23	3,24 5,29	3,01 4,77	2,85 4,44	2,74 4,20	2,66 4,03	2,59 3,89	2,54 3,78	2,49 3,69	2,45 3,61	2,42 3,55
	17	4,45 8,40	3,59 6,11	3,20 5,18	2,96 4,67	2,81 4,34	2,70 4,10	2,62 3,93	2,55 3,79	2,50 3,68	2,45 3,59	2,41 3,52	2,38 3,45

Продолжение таблицы 4

k_2		k_1 – степеней свободы числителя											
		1	2	3	4	5	6	7	8	9	10	11	12
k_2 - степени свободы знаменателя	18	4,41 8,28	3,55 6,01	3,16 5,09	2,93 4,58	2,77 4,25	2,66 4,01	2,58 3,85	2,51 3,71	2,46 3,60	2,41 3,51	2,37 3,44	2,34 3,37
	19	4,38 8,18	3,52 5,93	3,13 5,01	2,90 4,50	2,74 4,17	2,63 3,94	2,55 3,77	2,48 3,63	2,43 3,52	2,38 3,43	2,34 3,36	2,31 3,30
	20	4,35 8,10	3,49 5,85	3,10 4,94	2,87 4,43	2,71 4,10	2,60 4,87	2,52 3,71	2,45 3,56	2,40 3,45	2,35 3,37	2,31 3,30	2,28 3,23
	21	4,32 8,02	3,47 5,78	3,07 4,87	2,84 4,37	2,68 4,04	2,57 3,81	2,49 3,65	2,42 3,51	2,37 3,40	2,32 3,31	2,28 3,24	2,25 3,17
	22	4,30 7,94	3,44 5,72	3,05 4,82	2,82 4,31	2,66 3,99	2,55 3,76	2,47 3,59	2,40 3,45	2,35 3,35	2,30 3,26	2,26 3,18	2,23 3,12
	23	4,28 7,88	3,42 5,66	3,03 4,76	2,80 4,26	2,64 3,94	2,53 3,71	2,45 3,54	2,38 3,41	2,32 3,30	2,28 3,21	2,24 3,14	2,20 3,07
	24	4,26 7,82	3,40 5,61	3,01 4,72	2,78 4,22	2,62 3,90	2,51 3,67	2,43 3,50	2,36 3,36	2,30 3,25	2,26 3,17	2,22 3,09	2,18 3,03
	25	4,24 7,77	3,38 5,57	2,99 4,68	2,76 4,18	2,60 3,86	2,49 3,63	2,41 3,46	2,34 3,32	2,28 3,21	2,21 3,13	2,20 3,05	2,16 2,99
	26	4,22 7,72	3,37 5,53	2,98 4,64	2,74 4,14	2,59 3,82	2,47 3,59	2,39 3,42	2,32 3,29	2,27 3,17	2,22 3,09	2,18 3,02	2,15 2,96
	27	4,21 7,68	3,35 5,49	2,96 4,60	2,73 4,11	2,57 3,79	2,46 3,56	2,37 3,39	2,30 3,26	2,25 3,14	2,20 3,06	2,16 2,98	2,13 2,93
	28	4,20 7,64	3,34 5,45	2,95 4,57	2,71 4,07	2,56 3,76	2,44 3,53	2,36 3,36	2,29 3,23	2,24 3,11	2,19 3,03	2,15 2,95	2,12 2,90
	29	4,18 7,60	3,33 5,42	2,93 4,54	2,70 4,04	2,54 3,73	2,43 3,50	2,35 3,33	2,28 3,20	2,22 3,08	2,18 3,00	2,14 2,92	2,10 2,87
	30	4,17 7,56	3,32 5,39	2,92 4,51	2,69 4,02	2,53 3,70	2,42 3,47	2,34 3,30	2,27 3,17	2,21 3,06	2,16 2,98	2,12 2,90	2,09 2,84
	32	4,15 7,50	3,30 5,34	2,90 4,46	2,67 3,97	2,51 3,66	2,40 3,42	2,32 3,25	2,25 3,12	2,19 3,01	2,14 2,91	2,10 2,86	2,07 2,80
	34	4,13 7,44	3,28 5,29	2,88 4,42	2,65 3,93	2,49 3,61	2,38 3,38	2,30 3,21	2,23 3,08	2,17 2,97	2,12 2,89	2,08 2,82	2,05 2,76
	36	4,11 7,39	3,26 5,25	2,86 4,38	2,63 3,89	2,48 3,58	2,36 3,35	2,28 3,18	2,21 3,04	2,15 2,94	2,10 2,86	2,06 2,78	2,03 2,72
	38	4,10 7,35	3,25 5,21	2,85 4,34	2,65 3,86	2,46 3,54	2,35 3,32	2,26 3,15	2,19 3,02	2,14 2,91	2,09 2,82	2,05 2,75	2,02 2,69
	40	4,08 7,31	3,23 5,18	2,84 4,31	2,61 3,83	2,45 3,51	2,34 3,29	2,25 3,12	2,18 2,29	2,12 2,88	2,07 2,80	2,04 2,73	2,00 2,66

Продолжение таблицы 4

k_2		k_1 – степеней свободы числителя											
		1	2	3	4	5	6	7	8	9	10	11	12
k_2 - степени свободы знаменателя	42	4,07 7,27	3,22 5,15	2,83 4,29	2,59 3,80	2,44 3,49	2,32 3,26	2,24 3,10	2,17 2,96	2,11 2,86	2,06 2,77	2,02 2,70	1,99 2,64
	44	4,06 7,24	3,21 5,12	2,82 4,26	2,58 3,78	2,43 3,46	2,31 3,24	2,23 3,07	2,16 2,94	2,10 2,84	2,05 2,75	2,01 2,68	1,98 2,62
	46	4,05 7,21	3,20 5,10	2,81 4,24	2,57 3,76	2,42 3,44	2,30 3,22	2,22 3,05	2,14 2,92	2,09 2,82	2,04 2,73	2,00 2,66	1,97 2,60
	48	4,04 7,19	4,19 5,08	2,80 4,22	2,56 3,74	2,41 3,42	2,30 3,20	2,21 3,04	2,14 2,90	2,08 2,80	2,03 2,71	1,99 2,64	1,96 2,58
	50	4,03 7,17	3,18 5,06	2,79 4,20	2,56 3,72	2,40 3,41	2,29 3,18	2,20 3,02	2,13 2,88	2,07 2,78	2,02 2,70	1,98 2,62	1,95 2,56
	55	4,02 7,12	3,17 5,01	2,78 4,16	2,54 3,68	2,38 3,37	2,27 3,15	2,18 2,98	2,11 2,85	2,05 2,75	2,00 2,66	1,97 2,59	1,93 2,53
	60	4,00 7,08	3,15 4,98	2,76 4,12	2,52 3,65	2,37 3,34	2,25 3,12	2,17 2,95	2,10 2,82	2,04 2,72	1,99 2,63	1,95 2,56	1,92 2,50
	65	3,99 7,04	3,14 4,95	2,75 4,10	2,51 3,62	2,36 3,31	2,24 3,09	2,15 2,93	2,08 2,79	2,02 2,70	1,98 2,61	1,94 2,54	1,90 2,47
	70	3,98 7,01	3,13 4,92	2,74 4,08	2,50 3,60	2,35 3,29	2,23 3,07	2,14 2,91	2,07 2,77	2,01 2,67	1,97 2,59	1,93 2,51	1,89 2,45
	80	3,96 6,96	3,11 4,88	2,72 4,04	2,48 3,56	2,33 3,25	2,21 3,04	2,12 2,87	2,05 2,74	1,99 2,64	1,95 2,55	1,91 2,48	1,89 2,41
	100	3,94 6,90	3,09 4,82	2,70 3,98	2,46 3,51	2,30 3,20	2,19 2,99	2,10 2,82	2,03 2,69	1,97 2059	1,92 2,51	1,88 2,43	1,85 2,36
	125	3,92 6,84	3,07 4,78	2,68 3,94	2,44 3,47	2,29 3,17	2,17 2,95	2,08 2,79	2,01 2,65	1,95 2,56	1,90 2,47	1,86 2,40	1,83 2,23
	150	3,91 6,81	3,06 4,75	2,67 3,91	2,43 3,44	2,27 3,14	2,16 2,92	2,07 2,76	2,00 2,62	1,94 2,53	1,89 2,44	1,85 2,37	1,82 2,30
	200	3,89 6,76	3,04 4,71	2,65 3,88	2,41 3,41	2,26 3,11	2,14 2,90	2,05 2,73	1,98 2,60	1,92 2,50	1,87 2,41	1,83 2,34	1,80 2,28
	400	3,86 6,70	3,02 4,66	2,62 3,83	2,39 3,36	2,23 3,06	2,12 2,85	2,03 2,69	1,96 2,55	1,90 2,46	1,85 2,37	1,81 2,29	1,78 2,23
	1000	3,85 6,66	3,00 4,62	2,61 3,80	2,38 3,34	2,22 3,04	2,10 2,82	2,02 2,66	1,95 2,53	1,89 2,43	1,84 2,34	1,80 2,26	1,76 2,20
	∞	3,84 6,64	2,99 4,60	2,60 3,78	2,37 3,32	2,21 3,02	2,09 2,80	2,01 2,64	1,94 2,51	1,88 2,41	1,83 2,32	1,79 2,24	1,75 2,18

Продолжение таблицы 4

k_1 – степеней свободы числителя												k_1
14	16	20	24	30	40	50	75	100	200	500	∞	
245 6,142	146 6169	248 6208	249 6234	250 6258	251 6286	252 6302	253 6323	253 6334	254 6352	254 6361	254 6366	1
19,42 99,43	19,43 99,44	19,44 99,45	19,45 99,46	19,46 99,47	19,47 99,48	19,47 99,48	19,48 99,49	19,49 99,49	19,49 99,49	19,50 99,50	19,50 99,50	2
8,71 26,92	8,69 26,83	8,66 26,69	8,64 26,60	8,62 26,50	8,60 26,41	8,58 26,35	8,57 26,27	8,56 26,23	8,54 26,18	8,54 26,14	8,53 26,12	3
5,87 14,24	5,84 14,15	5,80 14,02	5,77 13,93	5,74 13,83	5,71 13,74	5,70 13,69	5,68 13,61	5,66 13,57	5,65 13,52	5,64 13,48	5,63 13,46	4
4,64 9,77	4,60 9,68	4,56 9,55	4,53 9,47	4,50 9,38	4,46 9,29	4,44 9,24	4,42 9,17	4,40 9,13	4,38 9,07	4,37 9,04	4,36 9,02	5
3,96 7,60	3,92 7,52	3,87 7,39	3,84 7,31	3,81 7,23	3,77 7,14	3,75 7,09	3,72 7,02	3,71 6,99	3,69 6,94	3,68 6,90	3,67 6,88	6
3,52 6,35	3,49 6,27	3,44 6,15	3,41 6,07	3,38 5,98	3,34 5,90	3,32 5,85	3,29 5,78	3,28 5,75	3,25 5,70	3,24 5,67	3,23 5,65	7
3,23 5,56	3,20 5,48	3,15 5,36	3,12 5,28	3,08 5,20	3,05 5,11	3,03 5,06	3,00 5,00	2,98 4,96	2,96 4,91	2,94 4,88	2,93 4,86	8
3,02 5,00	2,98 4,92	2,93 4,80	2,90 4,73	2,86 4,64	2,82 4,56	2,80 4,51	2,77 4,45	2,76 4,41	2,73 4,36	2,72 4,33	2,71 4,31	9
2,86 4,60	2,82 4,52	2,77 4,41	2,74 4,33	2,70 4,25	2,67 4,17	2,64 4,12	2,61 4,05	2,59 4,01	2,56 3,96	2,55 3,93	2,54 3,91	10
2,74 4,29	2,70 4,21	2,65 4,10	2,61 4,02	2,57 3,94	2,53 3,86	2,50 3,80	2,47 3,74	2,45 3,70	2,42 3,66	2,41 2,62	2,40 3,60	11
2,64 4,04	2,60 3,98	2,54 3,86	2,50 3,78	2,46 3,70	2,42 3,61	2,40 3,56	2,36 3,49	2,35 3,46	2,32 3,41	2,31 3,38	2,30 3,36	12
2,55 3,85	2,51 3,78	2,46 3,67	2,42 3,59	2,38 3,51	2,34 3,42	2,32 3,37	2,28 3,30	2,26 3,27	2,24 3,21	2,22 3,18	2,21 3,16	13
2,48 3,70	2,44 3,62	2,39 3,51	2,35 3,43	2,31 3,34	2,27 3,26	2,24 3,21	2,21 3,14	2,19 3,11	2,16 3,06	2,14 3,02	2,13 3,00	14
2,43 3,56	2,39 3,48	2,33 3,36	2,29 3,29	2,25 3,20	2,21 3,12	2,18 3,07	2,15 3,00	2,12 2,97	2,10 2,92	2,08 2,89	2,07 2,87	15
2,37 3,45	2,33 3,37	2,28 3,25	2,24 3,18	2,20 3,10	2,16 3,01	2,13 2,96	2,09 2,89	2,07 2,86	2,04 2,80	2,02 2,77	2,01 2,77	16
2,33 3,35	2,29 3,27	2,23 3,16	2,19 3,08	2,15 3,00	2,11 2,92	2,08 2,86	2,04 2,79	2,02 2,76	1,99 2,70	1,97 6,67	1,96 2,65	17

Продолжение таблицы 4

k_1 – степеней свободы числителя												k_1
14	16	20	24	30	40	50	75	100	200	500	∞	
2,29 3,27	2,25 3,19	2,19 3,07	2,15 3,00	2,11 2,91	2,07 2,83	2,04 2,78	2,00 2,71	1,98 2,68	1,95 2,62	1,93 2,59	1,92 2,57	18
2,26 3,19	2,21 3,12	2,15 3,00	2,11 2,92	2,07 2,84	2,02 2,76	2,00 2,70	1,96 2,63	1,94 2,60	1,91 2,54	1,90 2,51	1,88 2,49	19
2,23 3,13	2,18 3,05	2,12 2,94	2,08 2,86	2,04 2,77	1,99 2,69	1,96 2,63	1,92 2,56	1,90 2,53	1,87 2,47	1,85 2,44	1,84 2,42	20
2,20 3,07	2,15 2,99	2,09 2,88	2,05 2,80	2,00 2,72	1,96 2,63	1,93 2,58	1,89 2,51	1,87 2,47	1,84 2,42	1,82 2,38	1,81 2,36	21
2,18 3,02	2,13 2,94	2,07 2,83	2,03 2,75	1,98 2,67	1,93 2,58	1,91 2,53	1,87 2,46	1,84 2,42	1,81 2,37	1,80 2,33	1,78 2,31	22
2,14 2,97	2,10 2,89	2,04 2,78	2,00 2,70	1,96 2,62	1,91 2,53	1,88 2,48	1,84 2,41	1,82 2,37	1,79 2,32	1,77 2,28	1,76 2,26	23
2,13 2,93	2,09 2,85	2,02 2,74	1,98 2,66	1,94 2,58	1,89 2,49	1,86 2,44	1,82 2,36	1,80 2,33	1,76 2,27	1,74 2,23	1,73 2,21	24
2,11 2,89	2,06 2,81	2,00 2,70	1,96 2,62	1,92 2,54	1,87 2,45	1,84 2,40	1,80 2,32	1,77 2,29	1,74 2,23	1,72 2,19	1,71 2,17	25
2,10 2,86	2,05 2,77	1,99 2,66	1,95 2,58	1,90 2,50	1,85 2,41	1,82 2,36	1,78 2,28	1,76 2,25	1,72 2,19	1,70 2,15	1,69 2,13	26
2,08 2,83	2,03 2,74	1,97 2,63	1,93 2,55	1,88 2,47	1,84 2,38	1,80 2,33	1,76 2,25	1,74 2,21	1,71 2,16	1,68 2,12	1,67 2,10	27
2,06 2,80	2,02 2,71	1,96 2,60	1,91 2,52	1,87 2,44	1,81 2,35	1,78 2,30	1,75 2,22	1,72 2,18	1,69 2,13	1,67 2,09	1,65 2,06	28
2,05 2,77	2,00 2,68	1,94 2,57	1,90 2,49	1,85 2,41	1,80 2,32	1,77 2,27	1,73 2,19	1,71 2,15	1,68 2,10	1,65 2,06	1,64 2,03	29
2,04 2,74	1,99 2,66	1,93 2,55	1,89 2,47	1,84 2,38	1,79 2,29	1,76 2,24	1,72 2,16	1,69 2,13	1,66 2,07	1,64 2,03	1,62 2,01	30
2,02 2,70	1,97 2,62	1,91 2,51	1,86 2,42	1,82 2,34	1,76 2,25	1,74 2,20	1,69 2,12	1,67 2,08	1,64 2,02	1,61 1,98	1,59 1,96	32
2,00 2,66	1,95 2,58	1,89 2,47	1,84 2,38	1,80 2,30	1,74 2,21	1,71 2,15	1,67 2,08	1,64 2,04	1,61 1,98	1,59 1,94	1,57 1,91	34
1,98 2,62	1,93 2,54	1,87 2,43	1,82 2,35	1,78 2,26	1,72 2,17	1,69 2,12	1,65 2,04	1,62 2,00	1,59 1,94	1,56 1,90	1,55 1,87	36
1,96 2,59	1,92 2,51	1,85 2,40	1,80 2,32	1,76 2,22	1,71 2,14	1,67 2,08	1,63 2,00	1,60 1,97	1,57 1,90	1,54 1,86	1,53 1,84	38
1,95 2,56	1,90 2,49	1,84 2,37	1,79 2,29	1,74 2,20	1,69 2,11	1,66 2,05	1,61 1,97	1,59 1,94	1,55 1,88	1,53 1,85	1,51 1,81	40

Продолжение таблицы 4

k_1 – степеней свободы числителя												k_1
14	16	20	24	30	40	50	75	100	200	500	∞	
1,94 2,54	1,89 2,46	1,82 2,35	1,78 2,26	1,73 2,17	1,68 2,08	1,64 2,02	1,60 1,94	1,57 1,91	1,54 1,85	1,51 1,80	1,49 1,78	42
1,92 2,52	1,88 2,44	1,81 2,32	1,76 2,24	1,72 2,15	1,66 2,06	1,63 2,00	1,58 19,2	1,56 1,88	1,52 1,82	1,50 1,78	1,48 1,75	44
1,91 2,50	1,87 2,42	1,80 2,30	1,75 2,22	1,71 2,13	1,65 2,04	1,62 1,98	1,57 1,90	1,54 1,86	1,51 1,80	1,48 1,76	1,46 1,72	46
1,90 2,48	1,86 2,40	1,79 2,28	1,74 2,20	1,70 2,11	1,64 2,02	1,61 1,96	1,56 1,88	1,53 1,84	1,50 1,78	1,47 1,73	1,45 1,70	48
1,90 2,46	1,85 2,39	1,78 2,26	1,74 2,18	1,69 2,10	1,63 2,00	1,60 1,94	1,55 1,86	1,52 1,82	1,48 1,76	1,46 1,71	1,44 1,68	50
1,88 2,43	1,83 2,35	1,76 2,23	1,72 2,15	1,67 2,06	1,61 1,96	1,58 1,90	1,52 1,82	1,50 1,78	1,46 1,71	1,43 1,66	1,41 1,64	55
1,86 2,40	1,81 2,32	1,75 2,20	1,70 2,12	1,65 2,03	1,59 1,93	1,56 1,87	1,50 1,79	1,48 1,74	1,44 1,68	1,41 1,63	1,39 1,60	60
1,85 2,37	1,80 2,30	1,73 2,18	1,68 2,09	1,63 2,00	1,57 1,90	1,54 1,84	1,49 1,76	1,46 1,71	1,42 1,64	1,39 1,60	1,37 1,56	65
1,84 2,35	1,79 2,28	1,72 2,15	1,67 2,07	1,62 1,98	1,56 1,88	1,53 1,82	1,47 1,74	1,453 1,69	1,40 1,62	1,37 1,56	1,35 1,53	70
1,82 2,32	1,77 2,24	1,70 2,11	1,65 2,03	1,60 1,94	1,54 1,84	1,51 1,78	1,45 1,70	1,42 1,65	1,38 1,57	1,35 1,52	1,32 1,49	80
1,79 2,26	1,75 2,19	1,68 2,06	1,63 1,98	1,57 1,89	1,51 1,79	1,48 1,73	1,42 1,64	1,39 1,59	1,34 1,51	1,30 1,46	1,28 1,43	100
1,77 2,23	1,72 2,15	1,65 2,03	1,60 1,94	1,55 1,85	1,49 1,75	1,45 1,68	1,39 1,59	1,39 1,54	1,31 1,46	1,27 1,40	1,25 1,37	125
1,76 2,20	1,71 2,12	1,64 2,00	1,59 1,91	1,54 1,83	1,47 1,72	1,44 1,66	1,37 1,56	1,34 1,51	1,29 1,43	1,25 1,37	1,22 1,33	150
1,76 2,20	1,71 2,12	1,64 2,00	1,59 1,91	1,54 1,83	1,47 1,72	1,44 1,66	1,37 1,56	1,34 1,51	1,29 1,43	1,25 1,37	1,22 1,33	200
1,72 2,12	1,67 2,04	1,60 1,92	1,54 1,84	1,49 1,74	1,42 1,64	1,38 1,57	1,32 1,47	1,28 1,42	1,22 1,32	1,16 1,24	1,13 1,19	400
1,70 2,09	1,65 2,01	1,58 1,89	1,53 1,81	1,47 1,71	1,41 1,61	1,36 1,54	1,30 1,44	1,26 1,38	1,19 1,28	1,13 1,19	1,08 1,11	1000
1,69 2,07	1,64 1,99	1,57 1,87	1,52 1,79	1,46 1,69	1,40 1,59	1,35 1,52	1,28 1,41	1,24 1,36	1,17 1,25	1,11 1,15	1,00 1,09	∞

Таблица 5

Значение случайной величины равномерно распределенной в интервале (0.1)

0,10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
37542	04805	64894	74296	24805	24037	20363	10402	00822	91665
08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
66065	74717	34072	36850	36697	36170	35813	39855	11199	29170
31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
98520	17767	14905	68607	22109	40555	60970	93433	50500	73998
11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
99594	67348	87517	94969	91826	08928	93785	61368	23478	34113
65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
09893	20502	14225	68514	46427	56788	96297	78822	54382	14598
91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
80336	94598	26940	36858	70297	34135	53140	33340	42050	82348
44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
12550	73742	11100	02040	12860	74697	96644	89439	28707	25815
63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
61196	90446	26457	47774	51924	33729	65394	59593	42582	60527
15474	45266	95270	79953	59367	83848	82326	10118	33211	59466
94557	28573	67897	54387	54622	44431	91190	42592	92927	45973
42481	16213	97344	08721	16868	48767	03071	12059	25701	46670
23523	78317	73208	89837	68935	91416	26252	29663	05522	82562
04493	52494	75246	33824	45862	51025	61962	79335	65337	12472
00549	97654	64051	88159	96119	63896	54692	82391	23287	29529
35963	15307	26898	09354	33351	35462	77974	50024	90103	39333
59808	09391	45427	26842	83609	49700	13021	24892	78565	20106
46058	85236	01390	92286	77281	44077	93910	83647	70617	42941
32179	00597	87379	25241	05567	07007	86743	17157	85394	11838
69234	61406	20117	45204	15956	60000	18743	92423	97118	96338
19565	41430	01758	75379	40419	21585	66674	36806	84962	85207
45155	14938	19476	07246	43667	94543	59047	90033	20826	69541
94864	31994	36168	10851	34888	81553	01540	35456	05014	51176
98086	24826	45240	28404	44999	08896	39094	73407	35441	31880
33185	16232	41941	50949	89435	48581	88695	41994	37548	73043
80951	00406	96382	70774	20151	23387	25016	25298	94624	61171
79752	49140	71961	28296	69861	02591	71852	20539	00387	59579
18633	32537	98145	06571	31010	24674	05455	61427	77938	91936
74029	43902	77557	32270	97790	17119	52527	58021	80814	51748
54178	45611	80993	37143	05335	12969	56127	19255	36040	90324
11664	49883	52079	84827	59381	71539	09973	33440	88461	23356
48324	77928	31249	64710	02295	36870	32307	57546	15020	09994
69074	94138	86737	31976	35584	04401	10518	21615	01848	76938

Таблица 6

Квантили распределения Стьюдента $t^*_{1-\frac{\alpha}{2}}$

Число степеней свободы r	Уровень значимости α					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,17	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,74	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	4,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29

Таблица 7

Квантили распределения Фишера $F_{1-\alpha}$ для уровня значимости $\alpha = 0{,}05$

r_2	Число степеней свободы большей дисперсии r_1											
	1	2	3	4	5	10	20	30	40	60	120	
1	161,1	199,5	215,7	224,6	232,2	241,9	248,0	250,1	251,1	252,2	253,3	254,3
2	81,51	19,00	19,16	19,25	19,30	19,40	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,79	8,66	8,62	8,59	8,57	8,55	8,52
4	7,71	6,94	6,59	6,39	6,26	5,96	5,80	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,74	4,56	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,05	3,87	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,64	3,44	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,35	3,15	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,14	2,94	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	2,98	2,77	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	2,85	2,65	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	2,75	2,54	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,67	2,46	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,60	2,39	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,54	2,33	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,49	2,28	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,45	2,23	2,15	2,10	1,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,41	2,19	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,38	2,16	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,35	2,12	2,04	1,99	1,95	1,90	1,84
22	4,30	3,44	3,05	2,82	2,66	2,30	2,07	1,98	1,94	1,89	1,84	1,78
24	4,26	3,40	3,01	2,78	2,62	2,25	2,08	1,94	1,89	1,84	1,79	1,73
26	4,23	3,37	2,98	2,74	2,59	2,22	1,99	1,90	1,85	1,80	1,75	1,69
28	4,20	3,34	2,95	2,71	2,56	2,19	1,96	1,87	1,82	1,77	1,71	1,65
30	4,17	3,32	2,92	2,69	2,53	2,16	1,93	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,08	1,84	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	1,99	1,75	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	1,91	1,66	1,55	1,50	1,43	1,35	1,25
	3,84	3,00	2,60	2,37	2,21	1,83	1,57	1,46	1,39	1,32	1,22	1,00

Таблица 8

Преобразование Фишера $z = \frac{1}{2}\left[\ln(1 + r_{xy}) - \ln(1 - r_{xy})\right]$

r_{xy}	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,000	0,010	0,020	0,030	0,040	0,050	0,060	0,070	0,080	0,090
0,1	0,100	0,110	0,121	0,131	0,141	0,151	0,161	0,172	0,182	0,192
0,2	0,203	0,213	0,224	0,234	0,245	0,255	0,266	0,277	0,288	0,299
0,3	0,310	0,321	0,332	0,343	0,354	0,365	0,377	0,388	0,400	0,412
0,4	0,424	0,436	0,448	0,460	0,472	0,485	0,497	0,510	0,523	0,236
0,5	0,549	0,563	0,576	0,590	0,604	0,618	0,633	0,648	0,663	0,678
0,6	0,693	0,709	0,725	0,741	0,758	0,775	0,793	0,811	0,829	0,848
0,7	0,867	0,887	0,908	0,929	0,951	0,973	0,996	1,020	1,045	1,071
0,8	1,099	1,127	1,157	1,188	1,221	1,256	1,293	1,333	1,376	1,422
0,9	1,472	1,528	1,589	1,658	1,738	1,832	1,946	2,092	2,298	2,647

*) В дальнейшем, если это специально не оговорено, мы будем рассматривать только простые случайные выборки, которые для удобства будем называть случайными выборками или просто выборками.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Айвазян С. А., Бежаева З. Н., Староверов О. В. Классификация многомерных наблюдений. М.: Статистика, 1974, 240 с.

2. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Исследование зависимостей. М.: Финансы и статистика, 1985, 487с.

3. Айзерман М. А., Браверман Э. М., Розоноэр Л. И. метод потенциальных функций в теории обучения машин. М.: Наука, 1970, 384 с.

4. Беллман Р. Динамическое программирование. М.: Иностранная литература, 1960, 400с.

5. Боровков А. А. Математическая статистика. М.: Наука, 1984, 472 с.

6. Дуда Р., Харт П. Распознание образов и анализ сцен. М.: Мир, 1976, 512 с.

7. Де Гроот М. Оптимальные статистические решения. М.: Мир, 1974, 496 с.

8. Ивахненко А. Г., Юрачковский Ю. П. Моделирвоание сложных систем по экспериментальным данным. М.: Радио и связь, 1987, 120 с.

9. Казакевич Д. И. Основы теории случайных функций и ее применение в гидрометеорологии, 1977, 320 с.

10. Кендалл М. Дж., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973, 899 .

11. Кокс Д., Хинкли Д. Теоретическая статистика. М.: Мир, 1978, 560 с.

12. Себер Дж. Линейный регрессионный анализ. М.: 1980, 456 с.

13. Соболь И. М. Метод Монте-Карло. М.: Наука, 1978, 64 с.

14. Мудров В. И., Кушко В. Л. Методы обработки измерений. М.: Радио и связь, 1983, 304 с.

15. Соболь И. М., Статников Р. Б. Выбор оптимальных параметров в задачах с многими критериями. М.: Наука, 1981.

16. Справочник по прикладной статистике. В 2-х т. Пер. с англ. / под ред. Э Ллойда, У. Ледермана, Ю. Н. Тюрина. М.: Финансы и статистика, 1989.

17. Справочник по теории вероятностей и математической статистике. 2-е изд.. перераб. / Королюк В. С. и др. М.: Наука, 1985, 640 с.

18. Уланова Е. С., Забелин В. Н. Методы корреляционного и регрессивного анализа в агрометеорологии. Л.: Гидрометеоиздат, 1990, 208 с.

19. Шметтерер Л. Введение в математическую статистику. М.: Наука, 1976, 520 с.

20. Фомин Я. А., Тарловский Г. Р. Статистическая теория распознания образов. М.: Радио и связь, 1986, 264 с.

21. Хьюбер П. Робастность в статистике. М.: Мир, 1984, 304 с.

22. Юсупов Р. М. (ред.) Статистические методы в прикладной кибернетике. Изд-во МО СССР, 1980, 377 с.

23. Юсупов Р. М., Петухов Г. Б., Сидоров В. Н., Городецкий В. И., Марков В. М. Статистические методы обработки результатов наблюдений. МО СССР, 1984, 563 с.

24. Яглом А. М. Корреляционная теория стационарных случайных функций. Л.: Гидрометиздат, 1981, 280 с.

СОДЕРЖАНИЕ