МВД России
Санкт-Петербургский университет

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ ПРИ ПРОВЕДЕНИИ НАУЧНЫХ ИССЛЕДОВАНИЙ

Методические рекомендации

В 3-х частях

Часть 3

Санкт-Петербург
2023

План-график 2022, поз. 73

УДК 311
ББК 22.127
М33

М33 **Математико-статистические методы обработки экспериментальных данных при проведении научных исследований:** методические рекомендации: в 3-х частях. Часть 3 / Большакова Л.В., Сибаров К.Д., Яковлева Н.А. – Санкт-Петербург: СПбУ МВД России, 2023. – 92 с.

ISBN 978-5-91837-752-9
EDN: DAHTQH

В методических рекомендациях представлены методы многомерного математико-статистического анализа, позволяющие: разделить совокупность объектов на однородные группы; выделить главные факторы, характеризующие объекты; использовать экспертные мнения для математико-статистического анализа. Представлены основные понятия кластерного, дискриминантного и факторного анализа, метода экспертных оценок, а также рассмотрены основные методы анализа временных рядов. Для лучшего понимания материала в рекомендациях представлены примеры с решениями, приведены необходимые математико-статистические таблицы.

Предназначены для обучающихся в Санкт-Петербургском университете МВД России, занимающихся проведением научных исследований, в которых требуется исследовать взаимозависимость между характеристиками, признаками, переменными на основе статистических данных.

УДК 311
ББК 22.127

Рецензенты:
Меньших В. В., доктор физико-математических наук, профессор
(Воронежский институт МВД России);
Леонов А. П., кандидат юридических наук
(Уральский юридический институт МВД России)

ISBN 978-5-91837-752-9

Содержание

ВВЕДЕНИЕ

Предлагаемое издание является третьей частью методических рекомендаций, посвященных математико-статистическим методам обработки экспериментальных (выборочных) данных.

Как было отмечено в первой части рекомендаций математико-статистическое исследование какой-либо проблемы, связанное с выборочным методом, можно условно разделить на два этапа.

На первом этапе решаются задачи, связанные с исследованием одного количественного признака генеральной совокупности, а именно, сбор выборочных данных об этом признаке, представление этих данных в виде вариационных рядов, нахождение выборочных характеристик и, наконец, распространение выводов, полученных по выборочным данным, на всю генеральную совокупность, содержащую все значения признака.

Рассмотрению и решению вопросов и задач первого этапа была посвящена первая часть методических рекомендаций.

На втором этапе решаются задачи, связанные с исследованием двух и более количественных признаков генеральной совокупности, также с применением выборочного метода. Наиболее часто для решения таких задач применяют методы корреляционно-регрессионного анализа. С помощью этих методов можно с определенной вероятностью подтвердить или опровергнуть существование между двумя или несколькими признаками какой-либо зависимости. При условии существования этой зависимости оценить ее силу, найти приближенное функциональное представление, доказать значимость найденных коэффициентов и адекватность полученной модели зависимости одного признака от одного или нескольких других.

Рассмотрению и решению вопросов и задач корреляционно-регрессионного анализа была посвящена вторая часть методических рекомендаций.

В предлагаемой третьей части продолжается рассмотрение второго этапа математико-статистического исследования с применением многомерных методов.

Применение математико-статистических методов, как правило, предполагает что признаки, характеризующие элементы, являются количественными. Однако на практике бывает необходимо использовать и качественные признаки и, более того, иногда невозможно или очень сложно найти или оценить значения количественных призна-

ков. В этом случае для получения значений признаков можно использовать мнения экспертов, т. е. применить метод экспертных оценок, основные понятия которого рассмотрены в первой главе предлагаемых методических рекомендаций.

Как было отмечено выше, вторая часть рекомендаций посвящена понятиям и методам корреляционно-регрессионного анализа. Однако эти методы могут давать искаженные результаты, т. е. результаты, далекие от действительности, если не выполнены определенные условия для их применения. В частности, такая ситуация имеет место в том случае, когда рассматриваемая выборочная совокупность элементов является неоднородной, т. е. элементы существенно отличаются по значениям рассматриваемых признаков. При таком условии выборочную совокупность необходимо разбить на однородные группы, а затем внутри каждой группы проводить корреляционно-регрессионное исследование.

Проверку однородности совокупности элементов и, в случае неоднородности, разбиение их на однородные группы можно провести методами кластерного анализа. С кластерным анализом непосредственно связан дискриминантный анализ, с помощью которого можно оценить качество разбиения на однородные группы и выяснить, к какой группе можно отнести новый, добавленный в выборочную совокупность, объект. Основные понятия и методы дискриминантного анализа также рассмотрены во втором разделе.

В третьем разделе основное внимание уделено не объектам, а признакам, характеризующим эти объекты. Среди них, естественно, имеются наиболее и наименее важные для решения поставленной проблемы. Выделение главных факторов-признаков происходит с помощью методов факторного и компонентного анализов.

При решении ряда задач, связанных с выяснением взаимосвязей признаков, один из этих признаков может являться временным показателем. В этом случае математико-статистическая модель будет являться временным рядом. Исследованию временных рядов посвящен четвертый раздел.

Для лучшего понимания материала в пособии разобраны примеры с решениями. Приложения содержат необходимые математико-статистические таблицы.

1. МЕТОД ЭКСПЕРТНЫХ ОЦЕНОК

Современный этап развития общества ставит перед отдельными субъектами ряд задач, связанных с выбором дальнейших действий по повышению своих доходов, уменьшению риска и затрат, сохранению своей безопасности, улучшения здоровья и т. д. Для успешного решения подобных задач могут быть применены математико-статистические методы, и в частности методы теории принятия решений. С помощью методов этой теории могут быть получены способы оценки того или иного выбора действия, критерии надежности и безопасности этого действия, а также может быть сделан правильный прогноз относительно дальнейшего развития. Если при решении какого-то вопроса может быть получена полная математическая формализация рассматриваемой задачи, т. е. построена точная математическая модель, то для ее успешного решения достаточно выбрать и правильно применить соответствующий математико-статистический метод.

Однако довольно часто возникают ситуации, при которых невозможна полная математическая формализация проблемы, а значит, выбор, обоснование и оценка последствий не могут быть произведены на основе точных расчетов. В основном это происходит из-за сложности проблемы, недостатка информации, наличия не только количественных, но и качественных характеристик изучаемого явления в соответствующей предметной области, например, при решении задач по обеспечению экономической безопасности хозяйствующего субъекта, при решении ряда психологических или юридических проблем. В этом случае применяют метод экспертных оценок, в котором точные математические методы дополняются рекомендациями специалистов-экспертов, что позволяет хотя бы частично восполнить недостающую информацию, проверить ее достоверность или оценить качественные признаки.

Область применения метода достаточно широка. С помощью метода экспертных оценок могут быть решены, в частности, следующие задачи:

– установление степени предпочтительности (ранга) того или иного показателя, например, расположение факторов в порядке уменьшения степени их влияния на решение некоторой проблемы;

– определение оценок различных показателей, например, величины предполагаемых расходов для проведения какого-то мероприятия;

– определение оценки относительной важности показателей, объектов или критериев;

– нахождение альтернативных вариантов решения некоторой проблемы;

– определение наиболее вероятных интервалов времени свершения последовательности каких-то событий;

– решение задачи классификации, т. е. разбиение совокупности объектов на определенные группы по имеющимся признакам.

Особенностями данного метода, которые отличают его от обычной экспертизы, являются научная обоснованность организации проведения экспертного исследования и применение математико-статистических методов при обработке и анализе полученной экспертной информации. Эти особенности определяют основные этапы метода экспертных оценок.

1.1. Основные этапы метода экспертных оценок

Метод экспертных оценок состоит из следующих ряда этапов.

Этап 1. Формулировка проблемы, постановка задач.

Этап 2. Подготовка, организация и проведение экспертного оценивания.

Этап 3. Анализ и обработка экспертной информации.

Этап 4. Получение выводов, рекомендаций и прогнозов.

Для каждого из этапов существуют свои задачи, при решении которых могут применяться различные процедуры и методы, связанные с основной проблемой исследования.

Выполняемые на первом этапе формулировка проблемы, постановка целей и задач исследования имеют определяющее значение для метода экспертных оценок. Они, безусловно, связаны с потребностями практики и должны быть четко сформулированы руководителем данного исследования — лицом, принимающим решение (ЛПР). Для помощи в проведении данного исследования ЛПР может сформировать рабочую группу специалистов-аналитиков, в задачи которых входит организация и проведение экспертизы, а также, в некоторых случаях, обработка и анализ экспертной информации. При небольших исследованиях ЛПР может само организовать и провести экспертизу.

На втором этапе метода экспертных оценок решают вопросы, связанные с подготовкой, организацией и проведением экспертного оценивания. Необходимо отметить, что экспертное суждение не явля-

ется решением проблемы, это лишь полезная информация, которая помогает сделать правильный вывод. Одна из наиболее часто встречаемых ошибок экспертного оценивания состоит в том, что вначале собирают информацию, а потом решают, что с ней делать. В этом случае возможно либо получение излишней информации, либо, наоборот, ее нехватка.

На втором этапе должны быть проведены следующие мероприятия:

– формирование экспертной группы;
– выбор формы проведения процедуры экспертного оценивания;
– составление задания для экспертов и определение вида информации, которая должна быть от них получена;
– проведение процедуры экспертного оценивания и сбор экспертной информации.

Обработка и анализ информации, полученной от экспертов, составляет основное содержание третьего этапа.

Каждый эксперт представляет свою информацию в виде числовых данных, которые каким-то образом характеризуют определенные факторы, а также приводит содержательное обоснование этих данных. Наличие чисел и содержательных высказываний приводит к необходимости применения количественных и качественных методов обработки экспертной информации. Обработка и анализ информации может проводиться разными методами с использованием разных критериев, как правило, зависящих от формы проведения экспертизы и от вида экспертной информации. Однако есть ряд общих вопросов, решение которых составляет основное содержание третьего этапа метода экспертных оценок.

На третьем этапе должны быть решены следующие задачи:

– анализ индивидуальных оценок каждого эксперта;
– анализ совокупности оценок всей группы экспертов;
– объединение экспертных оценок.

После анализа и обработки экспертных данных переходят к последнему этапу метода экспертных оценок — 4 этапу, на котором формулируются выводы по рассматриваемой проблеме, принимаются какие-то решения, делаются возможные прогнозы.

1.2. Формирование экспертной группы

При формировании экспертной группы необходимо определить оптимальную её численность, получить правило оценки компетентности каждого эксперта и найти окончательный состав экспертной группы. Вопрос об оптимальном числе экспертов остается открытым, несмотря на то, что в некоторых работах предлагаются методы или формулы определения этого числа, но, как правило, без строгих обоснований. В большинстве таких формул используются понятие надежности или точности экспертизы, длины доверительного интервала оценок экспертов и т. п., то есть такие понятия, которые используются при формировании выборки в выборочном методе.

Однако количество экспертов не может определяться как объем выборки в выборочном методе на основе предельных теорем теории вероятностей, так как основное условие выборочного метода, по которому увеличение объема выборки приводит к увеличению точности результатов, для экспертного оценивания не выполняется. При создании экспертной группы необходимо учитывать, что при малом числе экспертов, может появиться излишнее влияние каждого отдельного эксперта, а при большом – возможна несогласованность и большой разброс во мнениях.

Практика показывает, что для получения качественной информации число экспертов не должно быть меньше пяти. С другой стороны, при числе экспертов больше двадцати пяти появляется излишняя информация. Кроме этого, заслуживают внимания соображения о существовании нижней и верхней границ числа экспертов. Нижняя граница равна числу факторов, оцениваемых в задаче, верхняя – определяется, например, как потенциальное число возможных экспертов. И, наконец, необходимо помнить о том, что наилучшего качества получаемой информации можно добиться подбором в экспертную группу наиболее компетентных экспертов-специалистов, а не увеличением их количества. В связи с этим особое внимание заслуживает задача определения компетентности экспертов. Для ее решения существует достаточно много методов, некоторые из которых дают очень неплохие результаты.

Однако лучше применять те методы, в которых степень компетентности определяется показателем, зависящим от самооценки эксперта и взаимооценки других специалистов. Заслуживает внимания также подход, при котором несколькими способами находят компе-

тентность некоторого эксперта, а при получении окончательной оценки его компетентности используют формулу

$$d_j = \frac{\sum_{i=1}^{n} a_{ij} p_i}{\sum_{i=1}^{n} p_i},$$

где a_{ij} — компетентность j-го эксперта, полученная i-ым способом; p_i — весовой коэффициент i-го способа.

Формирование состава экспертной группы, как правило, происходит в два этапа. Вначале определяется круг специалистов, компетентных в данной области, а затем из них выбирается экспертная группа с учетом оценки их компетентности.

Создание экспертной группы может происходить также по методу «снежного кома». Для применения этого метода необходимо, чтобы ЛПР или рабочая группа знали хотя бы нескольких специалистов, компетентных в рассматриваемой области. Этих специалистов просят назвать нескольких лиц, которые, на их взгляд, достаточно хорошо знают исследуемую проблему. Новых названных специалистов просят, в свою очередь, сделать то же самое, и т. д. Данный процесс может быть закончен тогда, когда уже имеющийся список не пополнится ни одним новым лицом. На практике, как правило, хватает двух-трех туров, после которых новые имена специалистов не появляются. После получения законченного списка специалистов начинают формировать из них экспертную группу с учетом компетентности экспертов и установленной численности группы.

1.3. Формы проведения процедуры экспертного исследования

Параллельно с формированием экспертной группы должна быть определена форма проведения процедуры экспертного оценивания. При анализе этих процедур необходимо ответить на два вопроса:

– каков должен быть характер взаимодействия экспертов в процессе опроса;

– должна ли существовать в процедуре оценивания обратная связь, т. е. возможно ли информирование экспертов о предыдущих турах опроса.

В зависимости от ответа на эти вопросы существуют четыре вида возможных процедур экспертного оценивания:

– одноразовые процедуры без обратной связи;

– одноразовые процедуры с обратной связью;

– многоразовые процедуры с обратной связью;

– многоразовые процедуры без обратной связи.

При экспертном оценивании по процедуре первого вида предполагается проведение одноразового раздельного опроса экспертов. Процедура проста в организации и не требует больших затрат на проведение. Как правило, ее применяют только в тех случаях, когда имеется дефицит времени и нет достаточно средств. Примерами такой процедуры могут служить обычное анкетирование или интервью.

Второй вид процедур предусматривает возможность общения экспертов во время экспертного оценивания. Примером могут служить различные совещания и дискуссии, на которых можно свободно отстаивать свою точку зрения, не допускается лишь критиковать ответы других. Наиболее распространенным способом получения экспертной информации по второй процедуре является метод «мозгового штурма». В основе этого метода лежит совместное обсуждение группой экспертов проблемной ситуации. Обсуждение должно удовлетворять следующим правилам:

– высказывания экспертов должны быть четкими и сжатыми;

– каждый эксперт может выступить несколько раз, но не подряд;

– замечания и критика предложений других экспертов не допускается;

– экспертам не разрешается зачитывать материал, подготовленный заранее, до обсуждения.

Дискуссии, которые проводятся в несколько четко выраженных туров, являются примером экспертного оценивания третьего вида. В ходе дискуссии эксперт может неоднократно высказывать свою точку зрения, менять ее.

Достоинства и недостатки для процедур второго и третьего вида практически одинаковы. Оперативная обратная связь, быстрое устранение недопонимания или необходимое уточнение формулировок – это, безусловно, достоинства этих процедур. К недостаткам относятся, прежде всего, влияние мнения авторитетного специалиста на мнения остальных, неготовность некоторых экспертов открыто признать свое мнение ошибочным, а также возможность возникновения «группового суждения», т. е. отказ экспертов от своих суждений во имя выработки единого мнения.

Основным методом, в котором используется процедура четвертого вида, является метод Дельфы. Этот метод часто применяется для решения задач прогнозирования, для оценки вероятности наступления тех или иных возможных в будущем событий. Он состоит из не-

скольких последовательно осуществляемых туров, результатом которых является формирование группового мнения по рассматриваемой проблеме. В первом туре эксперты отвечают на поставленные вопросы. Во втором туре каждый эксперт знакомится с ответами всех остальных, но без указания авторов ответов. Кроме этого эксперта знакомят с общим решением каждого тура, в результате чего каждый эксперт может поменять свое мнение или же аргументировать свой первоначальный ответ, если он существенно отличается от остальных. Третий тур — это повторение второго, только уже с новыми ответами (если они появились). После каждого тура специалисты-аналитики пытаются выработать единое усредненное решение и, когда это оказывается возможным, опрос заканчивается.

Для метода Дельфы обязательно выполнение трех условий: анонимность, существование регулируемой обратной связи и получение группового ответа.

Анонимность заключается в том, что в процессе проведения опроса эксперты не знают друг друга или, по крайней мере, не знают, какой ответ принадлежит какому эксперту. Анонимность позволяет исключить влияние отдельных достаточно известных экспертов на мнение остальных, а также при изменении мнения какого-то эксперта последний не обязан публично об этом объявлять.

Обратная связь позволяет устранить влияние индивидуальных и групповых интересов, не связанных с рассматриваемой проблемой. За последние годы появились различные модификации метода Дельфы, но суть этих методов остается та же.

1.4. Способы измерения и сравнения объектов

От задач и целей экспертного оценивания зависят вопросы и задания для экспертов, а также виды получаемой от них информации. Вопросы и задания должны быть четкими, понятными эксперту, в них не должно быть никаких наводящих на какой-либо ответ данных.

При формировании своих оценок эксперты могут использовать следующие способы измерения и сравнения объектов:

- ранжирование;
- парное сравнение;
- непосредственная оценка;
- последовательное сравнение.

Если для экспертов поставлена задача, в которой требуется как-то оценить факторы, не поддающиеся непосредственному измерению, то для ее решения может быть применен способ ранжирования. Ранжирование — это процедура упорядочения факторов, которая выполняется на основе предпочтения. С помощью ранжирования можно выбрать из всей совокупности факторов один или несколько наиболее существенных и важных по степени влияния на какой-то объект или явление. Процесс ранжирования наиболее важен и наиболее часто применяется для тех факторов, которые несоизмеримы в количественном отношении друг с другом. Можно выделить три главные задачи, в которых может быть успешно применен процесс ранжирования.

1. Имеется некоторое количество факторов (объектов, явлений), которые необходимо расположить во времени, например, выяснить в какой временной последовательности проводить их исследование. При этом в данной ситуации не требуется сравнение факторов по степени важности каких-то признаков, которыми они обладают.

2. Имеется некоторое количество факторов. Требуется упорядочить эти факторы по степени важности (предпочтительности) некоторого присущего им качественного признака или расположить эти факторы по степени их влияния на какой-то другой фактор, объект, явление. При этом точное измерение качественного признака не требуется.

3. Задача аналогичная предыдущей, только для факторов имеется не качественный, а количественный признак, но в настоящее время по каким-то причинам измерение его невозможно.

Суть способа ранжирования состоит в следующем. Каждый эксперт получает определенное количество факторов, которые он должен расположить в порядке некоторой очередности. При установлении очередности учитывается степень важности какого-то их признака или значимости влияния на какой-то объект. При этом самому предпочтительному фактору присваивается ранг (число), равный 1, следующему по предпочтению — ранг, равный 2, и т. д. Если эксперт для двух, трех или более факторов не может определить порядок их следования по предпочтительности, предположим, считает их одинаковыми по степени важности признака, то всем таким факторам приписывается одинаковый ранг, равный среднему суммы мест, поделенных между собой факторами с одинаковыми рангами. Например, если третий, четвертый и пятый факторы, по мнению эксперта, имеют

признак, одинаковый по важности, то каждый из этих факторов получает ранг, равный (3 + 4 + 5) : 3 = 4. Ранги, определенные таким образом, называют связанными. Значения связанных рангов могут быть и дробными.

Итак, в результате применения способа ранжирования от каждого эксперта будет получена информация в виде последовательности рангов, а также при необходимости обоснование поставленных оценок (рангов).

Главное достоинство способа ранжирования — его простота, а недостаток — в ограниченных возможностях использования. Точность и надежность данного способа существенно зависят от числа ранжируемых факторов. Чем меньше таких факторов, тем легче понять эксперту их различие и тем точнее и достовернее он может определить их ранги. Как правило, число факторов при применении этого способа не должно превышать двадцати, а для более надежных оценок — меньше десяти.

Если по каким-то причинам все же требуется провести упорядочение для большого числа факторов, то можно применять двухэтапный процесс ранжирования. Для этого все имеющиеся факторы разбивают на некоторое число групп факторов, схожих по смысловому содержанию. Задача экспертов — провести вначале ранжирование групп факторов, а затем отдельно в каждой группе — ранжирование факторов. В результате двухэтапного ранжирования от каждого эксперта будут получены ранги групп факторов и ранги факторов внутри каждой группы факторов. В чистом виде способ ранжирования редко используется при экспертном оценивании, чаще он применяется как часть более общего способа.

Трудности, возникающие при ранжировании большого числа факторов, значительно уменьшаются, если применять парные сравнения факторов. В способе парных сравнений эксперту из совокупности факторов предлагают поочередно всевозможные пары для предпочтительного выбора по какому-то признаку. Таким образом, здесь не требуется ранжировать всю совокупность факторов, а необходимо лишь из каждой пары выбрать фактор с наиболее предпочтительным признаком. Способ парного сравнения удобно использовать тогда, когда число рассматриваемых факторов достаточно велико или, когда различия между факторами настолько малы, что эксперту трудно провести обычное ранжирование. При применении парного сравне-

ния информация, полученная от каждого эксперта, может быть записана в виде таблицы.

Таблица 1.1.

Результаты оценок экспертов при проведении парного сравнения

Факторы	1	2	…	j	…	n
1	1	a_{12}	…	a_{1j}	…	a_{1n}
2	a_{21}	1	…	a_{2j}	…	a_{2n}
…	…	…	…	…	…	…
i	a_{i1}	a_{i2}	…	1	…	a_{in}
…	…	…	…	…	…	…
n	a_{n1}	a_{n2}	…	a_{nj}	…	1

Величины a_{ij} определяются по следующему правилу: $a_{ij} = 2$, если эксперт считает, что фактор i является более предпочтительным, чем фактор j; $a_{ij} = 1$, если факторы i и j имеют одинаковую степень важности; $a_{ij} = 0$, если фактор i является менее предпочтительным, чем фактор j. Способ парных сравнений может использоваться и для уточнения компетентности экспертов. Если в таблице, составленной экспертом, есть противоречия, то это может говорить о недостаточной компетентности данного эксперта.

Процедуры ранжирования и парного сравнения предполагают нахождение некой условной оценки — ранга фактора, по которой невозможно определить, насколько один фактор более значим, чем другой. Если для экспертов поставлена задача не только упорядочить факторы, но и найти вес каждого фактора в решении проблемы, то в процедуре экспертного оценивания применяется способ непосредственной оценки.

По этому способу эксперт должен поставить в соответствие каждому фактору некоторое число, характеризующее важность этого фактора. Числа-оценки определяются по одному из следующих двух способов. В первом способе предполагается известным или задается самим экспертом некоторый интервал, характеризующий диапазон изменения исследуемого качественного признака, присущего каждому фактору. Затем этот интервал разбивается на несколько интервалов, каждому из которых присваивается определенное числовое значение — оценка. Шкала оценок может быть своя у каждого эксперта и в ней могут быть как положительные, так и отрицательные значе-

ния. Задача эксперта состоит в том, чтобы каждый фактор поместить в определенный интервал в соответствии со значимостью присущего ему признака. Тем самым каждый фактор получает определенную оценку, соответствующую интервалу, в который он попадает.

Оценки факторов могут быть получены и по другому способу. Из всех факторов выбирается наиболее важный по оцениваемому признаку. Этому фактору присваивается оценка, равная некоторому фиксированному числу, например: 1, 10, 100. Остальным факторам присваиваются оценки, равные долям этого числа, в соответствии с важностью рассматриваемого признака. Итак, в результате применения способа непосредственной оценки от каждого эксперта должна быть получена информация об оценках каждого фактора.

Для нахождения оценок факторов может использоваться также метод последовательного сравнения. По этому способу эксперт должен проделать следующую работу:

– расположить все факторы в порядке уменьшения значимости признака;

– присвоить наиболее важному фактору оценку, равную p_1=1; остальным факторам – оценки p_2, p_3, …, p_n, равные долям единицы, в соответствии с важностью признака;

– сравнить значимость первого фактора с суммарной значимостью остальных, при этом оценки p_2, p_3, …, p_n остаются прежними, а значение p_1 при необходимости изменяется так, чтобы было выполнено неравенство $p_1 < p_2 + p_3 + \ldots + p_n$, если для эксперта первый фактор менее значим, чем сумма всех остальных, или выполнено неравенство $p_1 > p_2 + p_3 + \ldots + p_n$, в противном случае; если важность первого фактора совпадает с важностью остальных, то оценка первого фактора становится равной $p_1 = p_2 + p_3 + \ldots + p_n$;

– сравнить значимость второго фактора с суммарной значимостью всех остальных (без первого) и подкорректировать при необходимости оценку p_2 аналогично коррекции оценки p_1 в предыдущем пункте;

– продолжить процедуру до сравнения значимости (n–2)-го фактора с суммарной значимостью (n–1)-го и n-го факторов;

– разделить каждую вновь полученную оценку на сумму всех оценок, после чего получить окончательные нормированные оценки каждого фактора.

Достоинство данного способа состоит в том, что в процессе оценивания факторов эксперт сам анализирует свои оценки и исправ-

ляет, если это необходимо. Недостатки очевидны — это громоздкость и сложность. Требуется дополнительная работа с экспертами по обучению данной процедуре.

Приведенные способы сравнения и оценивания факторов могут использоваться экспертами и в комплексе, например, для упорядочения объектов может быть использовано парное сравнение и ранжирование с последующим анализом рангов факторов, полученных разными способами. После проведения всей вышеописанной подготовительной работы приступают к самой процедуре экспертного оценивания.

Организационно-финансовые проблемы подготовки и проведения экспертизы не являются предметом рассмотрения данной работы. Поэтому перейдем к третьему этапу метода экспертных оценок — к анализу и обработке экспертной информации.

1.5. Анализ оценок экспертов

После проведения процедуры экспертного оценивания и получения оценок экспертов необходимо провести анализ индивидуальных оценок каждого эксперта и затем анализ совокупности всех оценок на предмет их согласованности.

Каждый эксперт, даже достаточно компетентный, в силу различных причин, в том числе и случайного характера, может дать нелогичные или противоречивые ответы. Переход ко второй задаче данного этапа невозможен, если доказана противоречивость оценок эксперта. В этом случае необходимо провести соответствующие мероприятия, а иногда даже исключить эти оценки из дальнейшего рассмотрения.

Решение второй задачи предполагает анализ оценки всех экспертов на предмет их согласованности. При этом может возникнуть одна из следующих ситуаций. В первом случае может оказаться, что оценки различных экспертов похожи, достаточно близки друг другу, т. е. можно считать, что эти оценки группируются вблизи некоторого истинного значения. Таким образом, имеет место достаточно высокая общая согласованность экспертных оценок. Тогда для обработки результатов группового экспертного оценивания можно успешно применять методы математической статистики, основанные на осреднении данных, в результате чего появляется возможность выработать общее итоговое решение рассматриваемой проблемы.

В другом случае может оказаться, что все оценки экспертов разбиты на две или более групп похожих друг на друга оценок. Это может произойти, например, если в экспертной группе собраны представители разных течений или школ. В этом случае речь может идти не о выработке общего решения, а о получении нескольких выводов от каждой подгруппы экспертов с согласованными оценками. Необходим также качественный анализ причин возникновения таких группировок.

И, наконец, может возникнуть третья, самая неприятная ситуация, когда все или почти все оценки экспертов буду непохожими и далекими друг от друга. Причинами возникновения такой ситуации может быть в первую очередь плохой подбор экспертной группы, низкая компетентность экспертов.

Также на большой разброс оценок может повлиять неправильная или нечеткая формулировка вопросов, в результате чего может возникнуть разное понимание экспертами поставленной задачи. При возникновении третьей ситуации дальнейший анализ необходимо прекратить, далее рекомендуется уточнить постановку первоначальной проблемы, проверить компетентность экспертной группы и, возможно, пересмотреть ее состав, а также обратить внимание на более четкое разъяснение проблемы, вопросов и заданий экспертам.

Анализ согласованности оценок экспертов может быть произведен путем вычисления числовой меры, характеризующей степень близости индивидуальных мнений. Оценки экспертов могут быть в виде рангов и в виде произвольных чисел, характеризующих степень значимости каждого фактора для решения исследуемой проблемы. Если экспертная информация представлена в виде рангов, то согласованность оценок экспертов можно проверить по значению коэффициента конкордации ***W***.

Формула для нахождения этого коэффициента зависит от того, имеются ли хотя бы у одного эксперта связанные ранги, т. е. одинаковые ранги при оценке значимости разных факторов. Предположим вначале, что у всех экспертов все факторы имеют разные ранги, тогда коэффициент конкордации может быть вычислен по формуле:

$$W = \frac{12 \cdot S}{m^2 \cdot (n^3 - n)},$$

где m — число экспертов; n — число факторов; S — сумма квадратов отклонений суммы рангов каждого фактора, полученной от всех экс-

пертов, от среднего ранга. Например, процесс нахождения коэффициента конкордации для выяснения согласованности оценок восьми экспертов по ранжированию четырех факторов можно представить в виде таблицы.

Таблица 1.2

Расчет коэффициента конкордации W

Факторы, *n*	Эксперты, *m*								Сумма рангов	Отклонение суммы рангов	$(\Delta_i')^2$
	1	2	3	4	5	6	7	8			
	ранги								Δ_i	$\Delta_i' = \Delta_i - \overline{\Delta}$	
А	2	1	2	1	1	1	2	1	11	–9	81
Б	3	4	4	2	3	2	4	4	26	6	36
В	4	3	3	4	4	4	3	2	27	7	49
Г	1	2	1	3	2	3	1	3	16	–4	16
Общая сумма:									80		182
Средний ранг:	$\overline{\Delta}$ =80:4=20										

Отсюда:

$$W = \frac{12 \cdot 182}{8^2 \cdot (4^3 - 4)} \approx 0{,}57.$$

В случае, когда у одного или нескольких экспертов имеются одинаковые (связанные) ранги, коэффициент конкордации определяется по следующей формуле:

$$W = \frac{12 \cdot S}{m^2 \cdot (n^3 - n) - m \cdot \sum_{j=1}^{m} T_j},$$

Величина

$$T_j = \sum_{t_j} (t_j{}^3 - t_j),$$

где t_j — число одинаковых рангов, выставленных тем или иным экспертом.

Коэффициент конкордации измеряется в диапазоне от 0 до 1, причем, если $W = 0$, то имеет место полная несогласованность экспертов, если $W = 1$, то среди экспертов наблюдается полное единодушие. Наиболее реальным является случай частичной согласованности

оценок экспертов. На практике считают, что при $W > 0{,}5$ согласованность мнений экспертов существует.

Если коэффициент конкордации оказывается меньше 0,5, то согласованность между экспертами не подтверждается. В качестве причин этого могут быть, например, нечеткие постановка вопросов или инструктаж, неправильный выбор признаков, подбор некомпетентных экспертов, возможность сговора между ними и др. В этом случае рекомендуется поступить следующим образом:

– передать проведение ранжирования другой группе специалистов;

– изменить какие-то инструкции;

– скорректировать состав признаков;

– привлечь других экспертов.

При любом исходе проводить повторную экспертизу прежним составом экспертов не рекомендуется.

Если коэффициент конкордации больше 0,5, то можно провести проверку значимости коэффициента конкордации, т. е. проверить гипотезу о неслучайности согласованности экспертов. Проверку гипотезы можно провести по общей схеме, описанной в первой части методических рекомендаций.

Схема проверки данной гипотезы о значимости коэффициента конкордации содержит следующие пункты:

1. Основная гипотеза H_0: согласованность экспертов отсутствует;
альтернативная гипотеза H_1: согласованность экспертов существует.

2. Уровень значимости, чаще всего, полагают равным $\alpha = 0{,}05$.

3. Критерием является случайная величина, распределенная по закону Пирсона χ^2 (хи-квадрат).

4. Выборочное (расчетное) значение критерия по формуле:

$$\chi^2_{\text{в}} = W \cdot m \cdot (n - 1).$$

5. Критическая область является правосторонней. Критическая точка $\chi^2_{\text{кр}}$ находится по таблице распределения Пирсона χ^2 и зависит от уровня значимости α и числа степеней свободы $\vartheta = n - 1$.

6. Статистический вывод.

Если $\chi^2_{\text{в}} > \chi^2_{\text{кр}}$, то основная гипотез отвергается, коэффициент конкордации считается значимым, что свидетельствует о наличии

существенного сходства мнений экспертов и о неслучайности совпадений этих мнений.

Если $\chi^2_{\text{в}} < \chi^2_{\text{кр}}$, нет оснований отвергнуть основную гипотезу, т. е. считать значимым коэффициент конкордации. Также нет основания считать, что мнения экспертов являются согласованными. В этом случае вывод общих результатов на основании мнений экспертов будет не соответствовать действительности.

Проверим гипотезу о согласованности мнений экспертов для данных таблицы 1.2 при уровне значимости $\alpha = 0{,}05$.

Выборочное значение критерия равно

$$\chi^2_{\text{в}} = W \cdot m \cdot (n-1) = 0{,}57 \cdot 8 \cdot 3 = 13{,}68.$$

Значение критической точки, определенное по соответствующей таблице, в зависимости от $\alpha = 0{,}05$ и $\vartheta = n - 1 = 3$ равно

$$\chi^2_{\text{кр}} = 7{,}82.$$

Так как $\chi^2_{\text{в}} = 13{,}68 > \chi^2_{\text{кр}} = 7{,}82$, то согласованность мнений экспертов на уровне значимости $\alpha = 0{,}05$ подтверждена.

Как отмечалось ранее, оценки экспертов могут определять не ранги, а степени важности факторов, т. е. быть произвольными числами. В этом случае информацию, полученную от m экспертов с оценками важности n факторов, можно представить в виде таблицы.

Таблица 1.3

Результаты экспертизы по степени важности факторов

Эксперты	Факторы					
	1	2	…	j	…	n
1	a_{11}	a_{12}	…	a_{1j}	…	a_{1n}
2	a_{21}	a_{22}	…	a_{2j}	…	a_{2n}
…	…	…	…	…	…	…
i	a_{i1}	a_{i2}	…	a_{ij}	…	a_{in}
…	…	…	…	…	…	…
m	a_{m1}	a_{m2}	…	a_{mj}	…	a_{mn}

Эту таблицу можно преобразовать в вариационный статистический ряд с вариантами, равными значениям оценок a_{ij}. Тогда для ана-

лиза разброса оценок и их согласованности могут быть использованы известные статистические характеристики.

Наиболее часто для выяснения среднего разброса значений используют среднеквадратическое (стандартное) отклонение, определяемое по формуле

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{k}},$$

где x_i — варианты (оценки), $\bar{x}$ — средняя арифметическая оценок, k — число оценок.

При анализе согласованности оценок экспертов может быть также использован коэффициент вариации, который характеризует вариабельность оценок и определяется по формуле

$$V = \frac{\sigma}{\bar{x}} \cdot 100\%.$$

Кроме этого, для анализа разброса и согласованности оценок могут быть применены и другие специальные характеристики и показатели, подробно описанные в литературе по математической статистике и эконометрике.

1.6. Получение обобщенных результатов экспертного оценивания

Предположим теперь, что согласованность экспертов получена, тогда появляется задача объединения индивидуальных экспертных оценок в обобщенные групповые. В зависимости от целей экспертного оценивания, выбранного метода измерения и вида экспертной информации при обработке результатов возникает необходимость решения следующих основных задач:

1. Нахождение обобщенных рангов факторов на основе индивидуальных оценок экспертов, полученных после процедуры ранжирования.

2. Нахождение обобщенных рангов факторов на основе результатов парного сравнения факторов каждым экспертом.

3. Нахождение обобщенных относительных весов факторов на основе индивидуальных оценок экспертов.

4. Определение зависимостей между различными последовательностями рангов.

Решение первой задачи зависит от способа ранжирования факторов и вида информации, полученной от экспертов. Если для группы экспертов ставилась задача произвести обычное ранжирование имеющихся факторов, то для получения окончательных обобщенных рангов факторов поступают следующим образом. Вначале находят для каждого фактора сумму рангов, полученных от всех экспертов. Фактор, имеющий наименьшую сумму, получает окончательный ранг, равный 1, фактор, имеющий наименьшую сумму из оставшихся, получает ранг, равный 2, и т. д.

При применении двухэтапного процесса ранжирования информация от каждого эксперта представляется в виде двух последовательностей рангов: рангов групп факторов и рангов самих факторов внутри группы. В этом случае, задача получения обобщенных рангов делится на две части. В первой части необходимо преобразовать полученную от экспертов информацию в обычную последовательность рангов факторов. Во второй части по преобразованной информации — найти обобщенные ранги.

Очевидно, что решение второй части в точности повторяет вышеописанную процедуру получения обобщенных рангов при обычном ранжировании. Поэтому главной является первая часть общей задачи, которая состоит в преобразовании, полученной от каждого эксперта информации. Для такого преобразования может быть предложено несколько способов.

По первому способу ранг каждого фактора полагается равным номеру его места в последовательности факторов, сформированной по правилу: вначале располагаются факторы первой группы (группы, имеющей ранг 1) в порядке увеличения их рангов, затем присоединяют соответствующим образом факторы второй группы и т. д. В итоге, для каждого эксперта получают ранжирование всех имеющихся факторов. Очевидное достоинство такого способа — легкость и простота. Однако это ранжирование является достаточно грубым и может не учитывать существенные моменты. Например, по мнению эксперта, какой-нибудь фактор второй группы может иметь более важное значение, чем какой-то из факторов первой группы, однако по предложенному способу он получит ранг меньший, чем любой из факторов первой группы.

Устранить подобный недостаток можно при применении другого способа преобразования экспертной информации — ранжирования по сумме оценок. Для получения рангов факторов используют сравнение сумм двух оценок, а именно ранга группы, к которой относится фактор, и ранга самого фактора в группе. Полученное по этому способу ранжирование более объективно, так как учитывает степень различия в оценках мест групп и факторов внутри группы. Однако и здесь имеются недостатки, связанные, прежде всего, с тем, что число групп и число факторов в группах может не совпадать. Тогда приходится складывать величины, измеренные в разных масштабах. Для устранения этого недостатка может быть использован способ ранжирования по сравнимой шкале, в котором предполагается нахождение некоторого корректирующего коэффициента, при умножении на который можно определить оценки по сравнимой шкале.

При применении парного сравнения от каждого эксперта получают таблицу парных сравнений. Наиболее простой вариант определения последовательности рангов для каждого эксперта состоит в следующем. Находят сумму элементов каждой строки таблицы. Ранги факторов определяются по степени уменьшения этих сумм, причем ранг, равный 1, получает тот фактор, который расположен в строке с наибольшей суммой элементов.

После нахождения последовательности рангов факторов по таблице парных сравнений для каждого эксперта находят обобщенные ранги факторов по описанному выше способу.

Для нахождения обобщенных относительных весов факторов по индивидуальным оценкам экспертов требуется проведение определенных расчетов. Вначале происходит нормирование всех оценок для каждого эксперта, т. е. рассчитывается относительная значимость каждого фактора в отдельности. В качестве обобщенной оценки фактора принимают среднюю арифметическую из нормированных оценок этого фактора, полученных от всех экспертов.

Пример получения обобщенных относительных весов пяти факторов по индивидуальным оценкам двух экспертов приведен в таблице.

Таблица 1.4.

Результаты расчета обобщенных относительных весов пяти факторов по индивидуальным оценкам двух экспертов

Эксперты	Оценки факторов					Сумма оценок
	Ф1	Ф2	Ф3	Ф4	Ф5	
Э1	7	9	3	4	5	28
Э2	6	10	4	2	7	29
	Относительные оценки					
Э1	$\frac{7}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{4}{28}$	$\frac{5}{28}$	
Э2	$\frac{6}{29}$	$\frac{10}{29}$	$\frac{4}{29}$	$\frac{2}{29}$	$\frac{7}{29}$	
Сумма относительных оценок	0,457	0,666	0,245	0,212	0,42	
Обобщенные оценки	0,228	0,333	0,123	0,106	0,21	1

Важную роль при обработке и анализе экспертной информации играет задача установления зависимости между последовательностями рангов, полученных от двух экспертов или от двух групп экспертов. Предположим, что решается задача упорядочения факторов по степени их влияния на какой-то объект. Группы экспертов выбираются из двух разных научных школ, т. е. фактически происходит две процедуры экспертного оценивания для экспертов из одной и другой школы. В результате обработки экспертной информации получены две последовательности обобщенных рангов факторов, построенных на суждениях экспертов этих школ. Возникает вопрос: являются ли согласованными обобщенные мнения представителей двух направлений по рассматриваемому вопросу?

Для ответа могут быть использованы определенные коэффициенты, характеризующие степень ранговых зависимостей – коэффициенты Кендалла и Спирмена. Чаще всего зависимость между последовательностями рангов проверяется по критерию Спирмена. Для этого находят коэффициент ранговой корреляции по формуле

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} (x_i - y_i)^2}{n^3 - n},$$

где x_i и y_i ранги i-го фактора, поставленные представителями первой и второй школ соответственно. Этот коэффициент изменяется от –1 до

+1. Если величина коэффициента ρ близка к 1, то можно говорить о высокой согласованности мнений двух школ, если к 0, то согласованность мнений практически отсутствует. Если величина коэффициента близка к –1, то мнения представителей школ противоположны, т. е. факторы, имеющие наибольшие ранги по оценкам представителей одной школы, будут иметь наиболее низкие у представителей другой школы и наоборот.

Обработка результатов экспертного оценивания представляет собой достаточно трудоемкий процесс, особенно при большом числе факторов или экспертов. Поэтому для такой работы целесообразно применение средств вычислительной техники.

После анализа и обработки экспертных данных переходят к последнему этапу метода экспертных оценок — 4 этапу, на котором формулируются выводы по рассматриваемой проблеме, принимаются какие-то решения, делаются возможные прогнозы.

В качестве примера применения метода экспертных оценок для решения вопросов, связанных с экономической безопасностью, можно обратиться к научному исследованию в котором на основе метода экспертных оценок проведена оценка значимости факторов, влияющих на управленческое решение хозяйствующего субъекта о выборе вида взаимодействия с органами внутренних дел в системе противодействия теневым экономическим явлениям[1].

Были сформированы две группы экспертов. Одна группа состояла из руководителей хозяйствующих субъектов, другая — из сотрудников аппарата по борьбе с экономическими преступлениями. Число экспертов в каждой из групп было не менее тридцати. Экспертам требовалось провести ранжирование двадцати двух факторов. Так как число факторов оказалось достаточно большим, то применение процедуры обычного ранжирования было нецелесообразно. Процедура парного сравнения требовала больших временных и финансовых затрат. Поэтому лучше всего для ранжирования можно было бы применить любую модификацию метода непосредственной оценки, что и было проделано в работе.

---

[1] Ковтунова С. Ю. Механизм взаимодействия хозяйствующего субъекта с органами внутренних дел в системе противодействия теневым экономическим явлениям: автореф. дис. … канд. экон. наук. – СПб., 2011.

Ранжирование проводилось по следующему правилу. Для каждого фактора указывался диапазон изменения его признака, и эксперт должен был определить непосредственную оценку из этого интервала, которая, по его мнению, лучше всего характеризует рассмотренный признак.

Например, для фактора «Масштаб теневого сектора экономики» экспертам было предложено выбрать оценку из интервала от 1 до 10, причем выбор 1 означал отсутствие теневого сектора экономики, а 10 — все хозяйствующие субъекты, так или иначе, используют теневые практики ведения бизнеса.

Затем по каждому фактору была найдена сумма оценок всех экспертов группы. Фактору, получившему наименьшую сумму, присваивался ранг 1, соответственно, фактору с наибольшей суммой присваивался ранг 22. Полученные оценки позволили не только выявить наиболее влиятельные факторы с точки зрения представителей хозяйствующих субъектов и представителей органов внутренних дел, но и с помощь коэффициента Спирмена проверить согласованность мнений двух групп экспертов.

При сравнении обобщенных рангов, выставленных двумя экспертными группами, коэффициент Спирмена оказался достаточно близким к нулю, что позволило сделать следующий вывод: у представителей органов внутренних дел нет единого с представителями хозяйствующих субъектов взгляда на факторы, оказывающие влияние на выбор хозяйствующим субъектом вида взаимодействия с органами внутренних дел с точки зрения их важности и значимости. Естественно, такой вывод ставит перед исследователями дальнейшие задачи, например, получение ответов на вопросы:

1. Для каких факторов возникли наибольшие разногласия?
2. Почему возникли эти разногласия?
3. Существует ли возможность их устранения?

После применения метода экспертных оценпризнаки, характеризующие объекты, получают определенные количественные значения. Для дальнейшего исследования могут быть применены другие методы, и в частности методы корреляционно-регрессионного анализа. Однако перед применением этих методов совокупность объектов необходимо проверить на однородность с помощью кластерного анализа.

2. КЛАСТЕРНЫЙ И ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Как было отмечено во введении, методы корреляционно-регрессионного анализа могут давать результаты, далекие от действительности в случае, если выборочная совокупность состоит из неоднородных объектов, т. е. объектов, существенно отличающихся друг от друга по ряду характеристик. Предположим, например, что в качестве объектов выборки рассматриваются страны и ставится задача исследования взаимосвязи между определенными экономическими показателями, их характеризующими. Очевидно, что существование и сила взаимосвязи между показателями высокоразвитых стран и стран с низкой степенью развития экономики могут иметь существенные различия, иногда даже прямо противоположные. В этом случае, объединив эти страны в одну выборку, можно получить результаты, не соответствующие действительности.

Для того чтобы избежать выше описанных ошибок и получить результаты, соответствующие действительности, необходимо проверить выборку на однородность ее объектов и в случае подтверждения неоднородности, разделить всю совокупность на однородные группы.

Для разделения исследуемых совокупностей объектов, субъектов или явлений на однородные, в определенном смысле, группы используют методы кластерного и дискриминантного анализа. Само разделение на группы происходит с помощью кластерного анализа, дискриминантный анализ помогает проверить, уточнить или, при необходимости, подправить полученное распределение. После образования однородных групп может появиться новый объект, не принадлежащий ни к одной из полученных групп. С помощью методов дискриминантного анализа можно определить принадлежность этого объекта к одной из образованных групп.

Данный раздел посвящен основным понятиям и методам кластерного и дискриминантного анализа.

2.1. Понятие и сфера применения кластерного анализа

Первое применение кластерный анализ нашел в социологии. Его название происходит от английского слова *cluster* — гроздь, скопление. Предмет кластерного анализа и его первоначальное описание было сделано исследователем Трионом (Tryon) в 1939 году.

В начале 1960-х гг. кластерный анализ рассматривался как часть структурной лингвистики при «обучении» компьютеров человеческо-

му языку, которое состояло в том, чтобы научить распознавать буквы алфавитов (и других знаков) независимо от их начертания. Тогда же появились названия данной группы методов: «таксономия» и «распознавание образов». В «классической» статистике эти методы не существовали. Кроме этого появились термины «классификация с обучением» и «классификация без обучения».

Кластерный анализ предназначен для разделения (кластеризации) неоднородной совокупности, элементы которой характеризуются многими признаками, на однородные в определенном смысле или близкие по определяющим критериям группы (кластеры). Результатом его применения является объединение объектов в несколько групп, причём внутри каждой группы между объектами можно обнаружить сходство по заданным заранее критериям. Кластер это и есть группа.

Кластер-группы должны быть построены таким образом, чтобы внутри каждой группы оказались «похожие» объекты, а объекты разных группы должны существенно отличаться друг от друга.

Кластеризация и классификация — это не одно и то же. Главное отличие состоит в том, что при кластеризации перечень групп первоначально, как правило, не задан, он может определяться в процессе выполнения процедуры кластерного анализа. При классификации группы заранее описываются в соответствии с какими-то свойствами, затем каждый объект относят к заранее описанной группе.

Перед применением кластерного анализа предполагается, что у исследователя нет данных о том, каков должен быть их состав, каковы отличия кластеров друг от друга. Приступая к кластеризации, исследователю известны лишь объекты и признаки, их характеризующие. О сходстве или различии этих объектов судят по значениям этих признаков.

Результат применения кластерного анализа не связан с потерей какой-либо части информации, как, например, в случае использования факторного анализа, рассмотрению которого посвящен следующий раздел.

Итак, по своей сути кластеризация — это автоматическое разбиение данной совокупности элементов на некоторое число групп-кластеров в зависимости от схожести этих элементов.

Методы кластерного анализа могут применяться при решении многих задач.

Например, при изучении статистических данных, характеризующих какие-то объекты, разбиение множества этих объектов на группы позволяет лучше и более точно рассмотреть внутренние закономерности и взаимосвязи, на основании которых выдвинуть новые гипотезы о свойствах рассматриваемых объектов.

Кластеризация может рассматриваться в качестве подготовительного этапа перед применением корреляционно-регрессионного анализа, при этом очевидно, что дальнейшая обработка и построение общей модели значительно упрощается, так как каждый кластер вначале обрабатывается индивидуально.

Особое значение имеет кластеризации при решении проблемы сжатия информации. Если данные имеют большой объем, то после разбиения их на кластер-группы можно сократить объем данных, оставив по одному наиболее типичному представителю от каждого кластера или каждую группу считать за один общий объект, предварительно определив обобщающие характеристики признаков.

Кластеры используются не только для компактного представления объектов, но и для распознавания новых. В задачах прогнозирования можно предвидеть поведение объекта, к которому относится наблюдение, предположив, что оно будет схожим с поведением других объектов кластера.

И наконец, кластеризация применяется для выделения нетипичных (аномальных) наблюдений. Эту задачу также называют обнаружением аномалий. Интерес здесь представляют кластеры, в которые попадает крайне мало объектов, например, 1–3.

Методы кластерного анализа можно применять в самых различных случаях, даже в тех случаях, когда речь идет о простой группировке, в которой все сводится к образованию групп по количественному сходству. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не требует априорных предположений о наборе данных, не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет рассматривать множество исходных данных практически произвольной природы, а также анализировать показатели различных типов данных (интервальных данных, частот, бинарных данных). Отсутствие требования априорной информации объясняет популярность методов кластерного анализа в социальных науках, т. к. в большом количестве классификационных задач такая информация просто отсутствует.

Сфера использования кластерного анализа чрезвычайно обширна: тренеры ставят задачу выявления и отбора талантливых спортсменов; медики проводят кластеризацию заболеваний, лечения заболеваний или симптомов заболеваний; биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними; в археологии исследователи пытаются установить кластер-группы каменных орудий, похоронных объектов и т. д.; у экономистов возникает необходимость разбить исследуемые страны, банки, предприятия на приблизительно однородные группы в зависимости от различных экономических показателей, характеризующих эти объекты, чтобы провести анализ взаимосвязей экономических показателей внутри каждой группы; специалисты по обеспечению безопасности могут прийти к необходимости разделения всех исследуемых объектов на группы с приблизительно равным уровнем информационной или экономической безопасности, а затем внутри каждой группы провести исследование влияния и взаимосвязи различных признаков, определяющих уровень безопасности.

Важное значение кластерный анализ имеет применительно к совокупностям временных рядов, рассмотрению которых посвящен четвертый раздел. Здесь можно выделять периоды, когда значения соответствующих показателей были достаточно близкими, а также определять группы временных рядов, динамика которых наиболее схожа.

Перейдем к рассмотрению основных понятий кластерного анализа.

Пусть дана некоторая совокупность, состоящая из n объектов: Y_1, Y_2, …, Y_n. Каждый объект характеризуется m количественными признаками: P_1, P_2, …, P_m. Необходимо разделить имеющиеся объекты на однородные группы, используя значения признаков.

Очевидно, что, если $m = 1$, т. е. каждый объект характеризуется только одним признаком, то разделение объектов можно произвести по значениям этого признака, для чего не нужны никакие сложные методы анализа. При наличии двух признаков ($m = 2$) задача усложняется. Однако в этом случае провести кластеризацию можно, например, с применением графических представлений точек на плоскости.

Приведем простейший пример кластеризации объектов, которые характеризуются двумя признаками. Предположим, что первоначальная совокупность содержит 10 объектов, каждый их которых харак-

теризуется числовым значением двух признаков X и Y, представленных в таблице.

Таблица 2.1

Значения признаков объектов

Признак	Номер объекта									
	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
Признак P_1	1	7	9	8	2	8	3	1	7	2
Признак P_2	5	1	8	2	6	9	8	6	2	8

Отметим на плоскости 10 точек: на оси абсцисс фиксируем значения признака P_1, на оси ординат – P_2

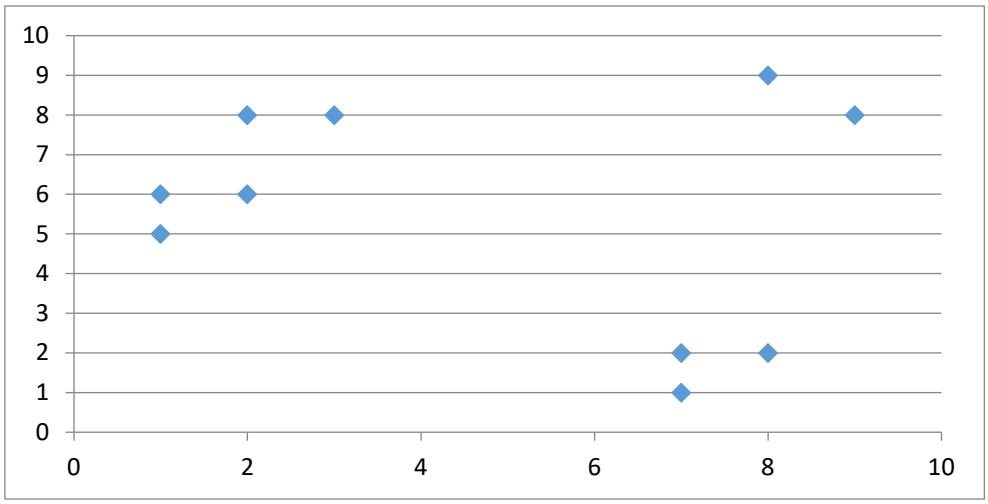

Рис. 2.1

На основании построенной диаграммы можно сделать вывод о том, что всю первоначальную совокупность можно разбить на три сравнительно однородные группы элементов: {1, 5, 7, 8, 10}; {2, 4, 9}; {3, 6}. После разбиения (кластеризации) десяти объектов, в зависимости от целей исследования можно, например, анализировать каждую группу в отдельности, или выбрать для анализа только одну группу, или каждую группу заменить на один элемент со средними значениями признаков и т. д.

Выделение трех групп в примере произошло чисто визуально, без какого-то научного обоснования, при этом существенным являлось то, что каждый элемент обладал только двумя признаками.

Естественно возникает вопрос, как провести кластеризацию в случае, когда у объектов фиксируются три и более признаков.

Кластерный анализ позволяет научно обоснованно провести кластеризацию элементов, характеризующихся не только двумя, но тремя и более числом признаков.

Методов кластерного анализа очень много, при этом, используя различные методы, аналитик может получить разные решения для одних и тех же данных. Однако чаще всего используют так называемые иерархические методы.

2.2. Иерархические методы кластерного анализа

Иерархические методы кластерного анализа применяют иерархические алгоритмы двух основных типов: восходящие (агломеративные, объединяющие, древовидные) и нисходящие (дивизимные, делимые) алгоритмы. Нисходящие алгоритмы работают по принципу «сверху вниз»: в начале все объекты помещаются в один кластер, который затем разбивается на все более мелкие кластеры. Более распространены восходящие алгоритмы, которые в начале работы помещают каждый объект в отдельный кластер, а затем объединяют кластеры во все более крупные, пока все объекты выборки не будут содержаться в одном кластере или число кластеров станет таким, каким требуется исследователю.

Таким образом, строится система вложенных разбиений. Результаты таких алгоритмов обычно представляют в виде дерева — дендрограммы (от греческого *dendron* – дерево). Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

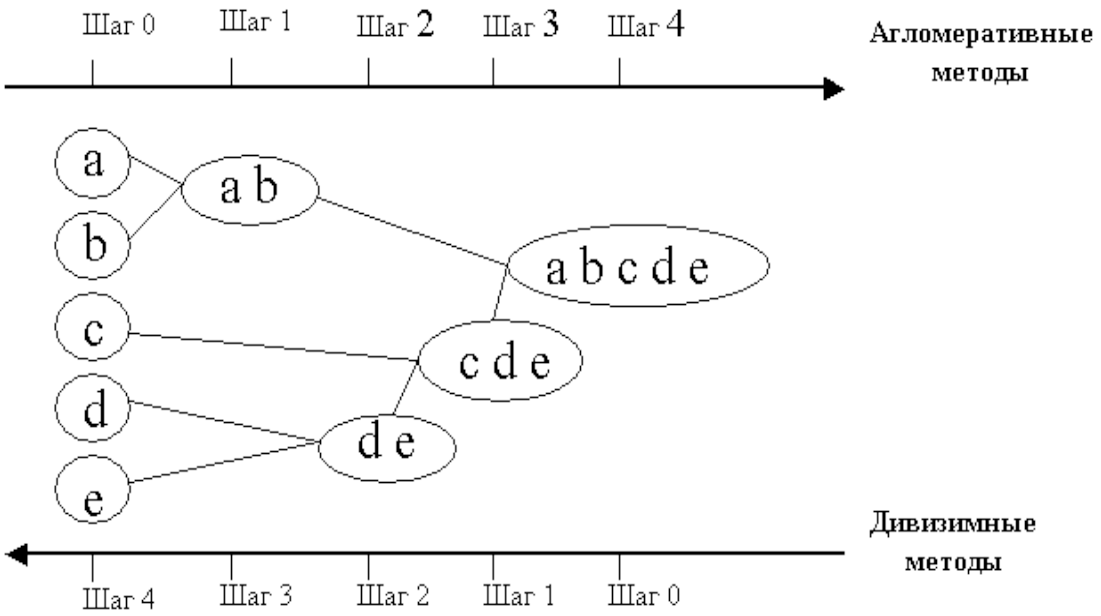

Рис. 2.2

Существует много способов построения дендрограмм. В дендрограмме объекты могут располагаться не только вертикально, но и горизонтально, например, рисунок 2.3.

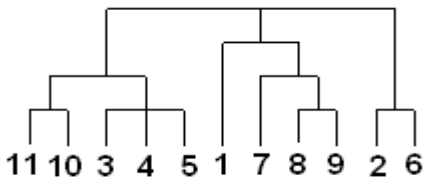

Рис. 2.3

Иерархические методы кластерного анализа чаще всего используются при небольших объемах наборов данных. Преимуществом иерархических методов кластеризации является их наглядность. К недостатку можно отнести систему полных разбиений, которая может являться излишней в контексте решаемой задачи.

Основными этапами иерархических методов кластерного анализа являются:

1. Отбор выборки объектов для кластеризации.

2. Выделение признаков, характеризующих каждый объект данной совокупности, получение их числовых значений. При необходимости — стандартизация значений признаков.

3. Определение меры близости между объектами и группами объектов.

4. Создание групп сходных объектов (кластеров).

5. Представление результатов кластеризации.

6. Анализ результатов кластеризации и при необходимости корректировка выбранной метрики (п. 3) и метода кластеризации (п. 4) до получения оптимального результата.

Первые два этапа, т. е. выбор объектов и признаков, их характеризующих, — это начальные этапы практически любого математико-статистического исследования. В зависимости от целей этого исследования объектами могут быть страны, предприятия, банки, люди и т. д. Каждый объект, как было отмечено выше, характеризуется определенным набором признаков.

Перед непосредственным применением метода кластерного анализа достаточно часто к признакам применяют процесс стандартизации. Необходимость применения этого процесса можно пояснить следующим образом. Предположим, например, что значения признака P_1 находятся в диапазоне от 100 до 800, а признака P_2 — в диапазоне от 0 до 1, т. е. значения признака P_1 на два порядка больше значений признака P_2. Тогда, при расчете расстояния между объектами, о чем речь пойдет далее, признак P_1 будет практически полностью доминировать над признаком P_2, значения которого значительно меньше. В этом случае, расстояние между объектами будет вычислено некорректно. Данная проблема может быть решена предварительной стандартизацией (*standardization*) значений признаков или нормированием (*normalization*). Эти процедуры приводят значения всех преобразованных признаков к единому диапазону значений. Чаще всего данные нормализуют вычитанием среднего и делением на стандартное отклонение. В результате такого процесса получают данные, для которых среднее значение оказывается равным нулю, а дисперсия оказывается равной единице.

На третьем этапе определяется метрика или мера близости между объектами и мера близости между группами объектов.

После выбора формул для нахождения расстояний между объектами и классами объектов можно переходить к непосредственному формированию групп, т. е. к самому процессу кластеризации.

На четвертом этапе происходит создание групп-кластеров с учетом выбора формулы расстояния между объектами и выбора правила нахождения расстояния между группами объектов.

На пятом этапе представляются кластер-группы, сам процесс кластеризации, а также необходимые промежуточные данные.

На шестом этапе происходит анализ полученных результатов, а также проверка достоверности разбиения на классы. Необходимо обратить внимание на то, что с помощью кластерного анализа разбить на группы можно любую совокупность объектов независимо от смыслового содержания как самих объектов, так и признаков, их характеризующих.

Рассмотрим основные этапы, начиная с третьего, более подробно.

Итак, пусть имеется некоторая совокупность, состоящая из n объектов: Y_1, Y_2, …, Y_n. Каждый объект характеризуется m количественными признаками: P_1, P_2, …, P_m. Эти объекты нужно разделить на однородные группы, используя значения признаков.

Все значения признаков представим в виде следующей матрицы:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix},$$

где x_{ij} — значение признака P_j для объекта Y_i.

На третьем этапе кластеризации исследователю приходится решать две задачи: задача определения расстояния между объектами и задача определения расстояния между кластерами (группами объектов).

Задача определения расстояния между объектами

Как было отмечено ранее, каждый из n объектов характеризуется m признаками, значениями которых являются столбцы матрицы X. По данным матрицы X формируется матрица расстояний D_1, размерности nxn, т. е. число строк и столбцов матрицы равно числу объектов:

$$D_1 = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{12} & d_{22} & \cdots & d_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ d_{1n} & d_{2n} & \cdots & d_{nn} \end{pmatrix}.$$

Элементами матрицы d_{ij} являются числа, определяющие некую меру близости (метрику) или расстояние между двумя объектами $d_{ij} = d(Y_i, Y_j)$.

Существует множество метрик, основными из которых являются следующие.

1. Евклидово расстояние:

$$d_{ij} = d(Y_i, Y_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}.$$

2. Квадрат Евклидова расстояния:

$$d_{ij} = d(Y_i, Y_j) = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2.$$

3. «Взвешенное» евклидово расстояние:

$$d_{ij} = d(Y_i, Y_j) = \sqrt{\sum_{k=1}^{m} w_k (x_{ik} - x_{jk})^2}.$$

4. Хеммингово расстояние:

$$d_{ij} = d(Y_i, Y_j) = \sum_{k=1}^{m} |x_{ik} - x_{jk}|.$$

5. Расстояние Чебышева:

$$d_{ij} = d(Y_i, Y_j) = \max_{1 \le k \le n} |x_{ik} - x_{jk}|.$$

Поясним кратко каждую из вышенаписанных формул.

Евклидово расстояние — это наиболее распространенная формула. Представляет собой геометрическое расстояние в многомерном пространстве. Использование этой формулы оправдано в случаях, когда признаки объектов имеют многомерное нормальное распределение; однородны по физическому смыслу и одинаково важны для классификации, признаковое пространство совпадает с геометрическим пространством.

Квадрат евклидова расстояния применяется для придания большего веса более отдаленным друг от друга признакам.

«Взвешенное» евклидово расстояние применяется в тех случаях, когда каждому признаку удается приписать некоторый вес, определяющий его значимость в общей исследуемой проблеме. Определение весов, как правило, связано с некоторыми дополнительными исследованиями. К этим исследованиям можно отнести, например, экс-

пертное оценивание признаков наиболее знающими эту проблему специалистами с дальнейшей математической обработкой этих мнений. Однако здесь очень важно правильно применить метод экспертных оценок. При неверном выборе весов, например, когда данные получены от некомпетентных экспертов, можно получить ложные выводы (см. первую главу данных методических рекомендаций).

Хеммингово (манхэттенское или расстояние городских кварталов) расстояние обычно применяется в качестве меры близости объектов, признаки которых являются дихотомическими. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (т. к. они не возводятся в квадрат).

Фактически расстояние Манхеттена — это кратчайшее расстояние между двумя точками, пройденное по линиям, параллельным осям прямоугольной системы координат. Само название «расстояние Манхеттена» возникло из-за ассоциаций, возникающих с прямоугольными формами застройки, которая характерна для современных городов. Преимущество расстояния Манхэттена заключается в том, что использование прямоугольной системы координат позволяет снизить влияние аномальных значений на работу алгоритмов кластеризации.

Расстояние Чебышева является максимумом модуля разности координат соответствующих признаков объектов. Это расстояние может оказаться полезным, когда необходимо определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением).

Выбор метрики полностью лежит на исследователе, поскольку результаты кластеризации могут существенно отличаться при использовании разных мер.

В процессе кластеризации возникает задача нахождения расстояния не только между объектами, но и между группами (кластерами) объектов.

Задача определения расстояния между кластер-группами

Существует несколько правил, по которым можно определить расстояние между двумя кластерами, хотя бы один из которых содержит два и более объекта. Например, в качестве расстояния между

кластер-группами можно считать расстояние между самыми близкими (или, наоборот, самыми далекими) объектами из разных групп.

Необходимо отметить, что перед задачей определения расстояния между кластерами должна быть решена предыдущая задача нахождения расстояния между объектами. Таким образом, для каждой пары объектов (Y_i, Y_j), где $Y_i \in S_1; Y_j \in S_2$ должно быть известно расстояние $d(Y_i, Y_j)$, найденное с использованием одной из описанных выше метрик.

Основными правилами, по которым определяются расстояния между кластерами, являются «правило ближайшего соседа», «правило дальнего соседа» и «правило средней связи». Опишем их более подробно.

Предположим, что требуется найти расстояние между двумя кластерами S_1 и S_2, каждый из этих кластеров состоит из определенного числа объектов. При этом один из этих кластеров может содержать и только один элемент.

1. Расстояние между кластерами по правилу «ближайшего соседа» (*Nearest Neighbor*) определяется по формуле:

$$d(S_1, S_2) = \min_{Y_i \in S_1; Y_j \in S_2} d(Y_i, Y_j).$$

Очевидно, что по этому правилу расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах.

Это правило иногда называют методом или правилом одиночной связи (*Single Linkage*). Этот метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов. В результате работы этого метода кластеры представляются длинными «цепочками» или «волокнистыми» кластерами, «сцепленными вместе» только отдельными элементами, которые случайно оказались ближе остальных друг к другу. Тенденция данного метода сводится к образованию небольшого числа крупных кластеров.

Как альтернативу можно использовать соседей в кластерах, которые находятся дальше всех остальных пар объектов друг от друга.

2. Расстояние между кластерами по правилу «дальнего соседа» (*Furthest Neighbor*) определяется по формуле:

$$d(S_1, S_2) = \max_{Y_i \in S_1; Y_j \in S_2} d(Y_i, Y_j).$$

Это правило иногда называют методом полной связи (*Complete Linkage*). В этом методе расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т. е. наиболее удаленными соседями). Тенденция данного метода сводится к образованию большего числа компактных кластеров.

Этот метод обычно работает очень хорошо, когда объекты на самом деле происходят из реально различных групп. Если же кластеры имеют удлиненную форму или их естественный тип является «цепочечным», то этот метод непригоден.

3. Расстояние между кластерами по правилу «средней связи» (*Average Linkage*) или межгрупповой связи определяется, как среднее арифметическое всех попарных расстояний между объектами рассматриваемых групп, т. е. по формуле:

$$d(S_1, S_2) = \frac{1}{n_1 \cdot n_2} \sum d(Y_i, Y_j),$$

где сумма берется по всем парам объектов (Y_i, Y_j), $Y_i \in S_1; Y_j \in S_2$, величины n_1 и n_2 определяют число объектов в первой и второй группах соответственно.

Иногда это правило называют методом невзвешенной попарной средней. Этот метод занимает промежуточное положение между первыми двумя вышеописанными методами, поэтому он должен давать более точные результаты, чем предыдущие. Объединение в кластеры при данном методе происходит при расстоянии большем, чем в методе одиночной связи, но меньшем, чем при использовании метода полной связи.

Несмотря на то, что метод средней связи дает более точные результаты, остальные два метода также применяются в реальных исследованиях. Например, при «сжатии» пространства, т. е. образовании минимально возможного числа крупных кластеров, применяют метод одиночной связи, при «расширении» пространства, т. е. при получении максимально возможного числа компактных кластеров — метод полной связи.

Кроме вышеописанных трех методов, иногда применяют метод взвешенного попарного среднего, который идентичен предыдущему

методу, за исключением того, что при вычислениях размер соответствующих кластеров (т. е. число объектов, содержащихся в них) используется в качестве весового коэффициента. Поэтому данный метод должен быть использован, когда предполагаются неравные размеры кластеров.

В невзвешенном центроидном методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести. Идентичен ему взвешенный центроидный метод (метод медиан), в котором при вычислениях используются веса для учета разницы между размерами кластеров. Поэтому, если имеются или подозреваются значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.

И наконец, для нахождения расстояния между кластерами может использоваться метод Варда (*Ward's method*). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения (Ward, 1963). В отличие от других методов кластерного анализа для оценки расстояний между кластерами, здесь используются методы дисперсионного анализа. На каждом шаге кластеризации будут объединяться такие два кластера, которые приводят к минимальному увеличению целевой функции, т. е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров и «стремится» создавать кластеры малого размера.

После выбора формул для нахождения расстояний между объектами и классами объектов можно переходить к четвертому этапу процесса кластеризации, т. е. непосредственному формированию групп.

Методы создания групп-кластеров

Рассмотрим для примера более подробно иерархический восходящий алгоритм кластеризации.

Предположим, что формулы расстояния между объектами и между группами объектов выбраны.

Итак, вначале предполагается, что каждый элемент совокупности — это отдельный кластер и между кластерами по какой-то из описанных выше формул найдено расстояние в виде матрицы D_1, симметричной относительно главной диагонали:

$$D_1 = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{12} & d_{22} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{1n} & d_{2n} & \dots & d_{nn} \end{pmatrix},$$

где $d_{ij} = d(Y_i, Y_j)$.

Из всех расстояний выбираем наименьшее и объединяем соответствующие этому расстоянию элементы в одну группу (кластер). В результате два (или более) столбца и две (или более) строки матрицы D_1 преобразуются в один. Пересчитываем на основе одного из правил нахождения расстояния между вновь образованным кластером и остальными элементами. Получаем новую матрицу расстояний D_2, в которой будет, по крайней мере, на одну строку и один столбец меньше по сравнению с D_1. Далее в матрице D_2 находим минимальное расстояние и аналогично объединяем соответствующие кластеры и т. д. Процесс может быть закончен на любой стадии в зависимости от того, какое количество групп требуется в задаче. Кроме этого, на каждом шаге объединения кластеров фиксируем значение минимального расстояния t_i, $i = 1, \dots$. Эти значения могут использоваться при определении числа кластеров. Процедуре группировки в иерархическом кластерном методе соответствует постепенное скачкообразное увеличение чисел $t_1, t_2, \dots, t_i, \dots$ Резкое увеличение значения t_i можно считать характеристикой числа кластеров, которые реально существуют в данной совокупности объектов. Таким образом, резко измененное значение величины t_i характеризует переход от сильно связанного состояния объектов к слабо связанному.

2.3. Краткий обзор методов кластерного анализа

Кроме иерархических существуют и другие методы кластеризации. Рассмотрим некоторые из них.

Предположим, что для исследователя важно, чтобы после кластеризации было образовано ровно k групп. Таким образом, необходимо применить такие методы кластерного анализа, в результате применения которых будет образовано ровно k кластеров, но так, чтобы они были настолько различны, насколько это возможно. Такого типа задачи могут быть решены методом k-средних. В общем случае алгоритм метода k-средних построит ровно k кластер-групп, которые будут расположены на самых больших по возможности расстояниях друг от друга.

Сам алгоритм метода *k*-средних начинается со случайного выбора *k* объектов-кластеров. Затем начинается присоединение к кластерам остальных объектов, которые не принадлежат выбранным, а далее изменение принадлежности объектов к кластерам так, чтобы, во-первых, изменчивость внутри кластеров минимизировать, а между кластерами – максимизировать. Этот метод иногда называют методом «дисперсионного анализа наоборот» в следующем смысле. При проверке гипотезы о различии средних в группах друг от друга в дисперсионном анализе сравнивают межгрупповую дисперсию (изменчивость) с внутригрупповой. При применении метода *k*-средних при кластеризации происходит перемещение объектов из одних групп в другие с целью получить наиболее значимый результат с точки зрения дисперсионного анализа.

Все алгоритмы методов кластерного анализа можно условно разделить на четкие и нечеткие. В четких алгоритмах каждый объект принадлежит только одному кластеру, т. е. каждому объекту ставится в соответствие номер кластера. При нечетких алгоритмах каждый объект относится к определенному кластеру с некоторой вероятностью и, таким образом, каждому объекту ставится в соответствие некоторое число, которое показывает степень отношения данного объекта к соответствующему кластеру.

Наиболее часто применяемым алгоритмом нечеткой кластеризации является алгоритм *c*-средних (*c*-means). Его можно рассматривать, как некую модификацию метода *k*-средних.

Существуют алгоритмы кластеризации, которые основаны на теории графов. В таких алгоритмах выборка представляется в виде графа. Вершины графа — это объекты, а ребра имеют определенный вес, численно равный расстоянию между объектами. Эти алгоритмы обладают рядом преимуществ, а именно — они более наглядны, относительно просты в реализации и существует возможность, основываясь на геометрических соображениях, внести различные усовершенствования. Основными алгоритмами такого вида являются алгоритм выделения связных компонент, алгоритм построения минимального покрывающего (остовного) дерева и алгоритм послойной кластеризации.

До настоящего момента была рассмотрена задача кластеризации объектов. Однако кластеризацию можно проводить и для признаков, получая при этом достаточно интересные результаты. Более того, можно проводить кластеризация в обоих направлениях и по объектам,

и по признакам. Это позволяет сделать так называемая двухвходовая процедура. Используется эта процедура достаточно редко и в тех случаях, когда и объекты, и признаки вносят существенный вклад в создание осмысленных кластеров.

Программная реализация алгоритмов кластерного анализа широко представлена в различных инструментах Data Mining, которые позволяют решать задачи достаточно большой размерности. Например, агломеративные методы реализованы в пакете SPSS, дивизимные методы — в пакете Statgraf.

В заключение важно отметить, что методы кластерного анализа применяются только для тех объектов, которые по смыслу допускают разбиение их по группам в соответствии со значениями признаков, характеризующих эти объекты.

С кластерным анализом тесно связан дискриминантный анализ, который позволяет, во-первых, отнести вновь появившийся объект к одной из образованных кластер-групп и, во-вторых, проверить и, при необходимости, уточнить правильность кластеризации.

2.4. Понятие и назначение дискриминантного анализа

Предположим, что после проведения кластерного анализа, т. е. получения однородных групп, было выяснено, что появились новые объекты, которые обязательно должны быть включены в исследование. Возникает вопрос, к какой группе отнести новые объекты? Ответ на этот и ряд других вопросов, связанных с классификацией, дает дискриминантный анализ.

Дискриминантный анализ (*discriminant analysis*) — метод многомерного статистического анализа. Он включает в себя методы классификации многомерных наблюдений по принципу максимального сходства с элементами ранее исследованных групп. Иначе говоря, дискриминантный анализ позволяет определить принадлежность новых объектов к образованным группам.

Исходными данными для дискриминантного анализа является множество объектов, разделенных на группы так, что каждый объект может быть отнесен только к одной группе. Для каждого из объектов имеются данные по ряду количественных признаков, из которых в дискриминантном анализе образуются новые переменные, называемые дискриминантными переменными или предикторами.

Задачами дискриминантного анализа является определение:

– решающих правил, позволяющих по значениям дискриминантных переменных (предикторов) отнести каждый объект к одной из известных групп;

– «веса» каждой дискриминантной переменной для разделения объектов на группы.

В отличие от кластерного анализа новые кластеры не образуются, а являются правилом, по которому объекты относятся к определенной группе. Задача состоит в том, чтобы вновь поступающий объект отнести в одну из имеющихся групп. У понятия «дискриминация» имеется много синонимов: диагностика, распознавание образов, математическая и статистическая классификация и т. д.

Предположим, что имеется множество объектов, заданных значениями признаков и принадлежность которых к той или иной группе (кластеру) достоверно известна. Смысл дискриминантного анализа – на основании имеющегося множества объектов преобразовать многомерный массив в одномерный показатель для прогнозирования принадлежности новых объектов к группам, т. е. построить новый обобщенный показатель, значения которого максимально различаются для объектов, отнесенных к разным группам.

Исследование различий между группами — основа концепции дискриминантного анализа. Например, исследуется принадлежность некоторого товара к группам товаров марок А и В. При проведении дискриминантного анализа находят дискриминантную функцию (линейную комбинацию характеристик товаров групп А и В), которая наилучшим образом различает категории или группы. Затем по значению этой функции для нового товара определяют, какую марку ему можно присвоить.

В маркетинге, например, можно получить ответы на вопросы: какими демографическими характеристиками обладают приверженцы бренда, как различаются между собой сегменты рынка, какими характеристиками обладают потребители, реагирующие на прямую почтовую рассылку и др. Для оценки финансового состояния своих клиентов при выдаче им кредита банк может классифицировать их по надежности на несколько категорий по ряду признаков. В случае, когда следует отнести клиента к той или иной категории, используют процедуры дискриминантного анализа.

Использование методов дискриминантного анализа помогает ответить на вопросы из различных областей практики. Например, подходит ли соискатель работы для той или иной должности? Отли-

чаются ли в потреблении замороженных продуктов покупатели, которые пьют безалкогольные напитки мало, умеренно и много? Какие факторы влияют на увеличение риска пациента получить сердечный приступ? Какие психографические характеристики помогают провести различия между восприимчивыми и не восприимчивыми к цене покупателями компьютерной техники? Различаются ли между собой сегменты рынка по своим предпочтениям к средствам массовой информации?

Дискриминантный анализ применяется также в теории распознавания образов, развивающей теоретические основы и методы классификации и идентификации предметов, явлений, процессов, ситуаций и прочих объектов, которые характеризуются конечным набором свойств и признаков. Распознавание образов — это, по сути, отнесение исходных данных к определенному классу с помощью выделения существенных признаков, характеризующих эти данные, из общей массы несущественных данных.

Дискриминантный анализ можно отнести к методам анализа зависимости, при этом внешний вид получаемой дискриминантной функции не будет отличаться от уравнения линейной множественной регрессии:

$$U = a_0 + a_1 z_1 + a_2 z_2 + \ldots + a_k z_k.$$

В качестве зависимой переменной выступает номинальная переменная, идентифицирующая принадлежность объектов к одной из нескольких групп. Независимые переменные z_1, z_2, …, z_k такие же, как и в регрессионном анализе (количественные и качественные); неизвестные коэффициенты a_1, a_2, …, a_k должны быть оценены в результате применения дискриминантного анализа. Максимально точно подобрать значения неизвестных коэффициентов помогает то, что уравнение регрессии составляется на основе тех объектов, о которых известна групповая принадлежность.

Таким образом, по внешним признакам модель дискриминантного анализа похожа на модель логистической регрессии, но эти модели отличаются способами вычисления коэффициентов.

Коэффициенты a_1, a_2, …, a_k должны быть определены так, чтобы по значениям дискриминантной функции можно было с максимальной четкостью провести разделение по группам, т. е. группы максимально возможно должны отличаться значениями дискриминантной функции. Это происходит тогда, когда отношение межгрупповой суммы квадра-

тов к внутригрупповой сумме квадратов для дискриминантных показателей максимально. Затем на основе этого же правила отнести (предсказать) новый объект или множество новых объектов к «свойственной» им группе.

Задачей дискриминантного анализа является исследование групповых различий — различение (дискриминация) объектов по определенным признакам. Дискриминантный анализ позволяет выяснить, действительно ли группы различаются между собой, и если да, то каким образом, т. е. какие переменные вносят наибольший вклад в имеющиеся различия.

Существует два основных вида дискриминантного анализа:

– дискриминантный анализ для двух групп (*two-group discriminant analsysis*) — зависимая переменная имеет две категории.

– множественный дискриминантный анализ (*multiple descriminant analysis*) — у зависимой переменной имеется три или больше категорий.

При сравнении двух групп (бинарная зависимая переменная) формируется одна дискриминантная функция. Если данный метод применяется к анализу трех или более групп (множественный дискриминантный анализ), то могут формироваться несколько дискриминантных функций.

Дискриминантный анализ имеет определенное сходство с кластерным анализом. Целью того и другого анализа является разделение совокупности объектов на несколько групп. Однако сам процесс разделения в двух видах анализа принципиально различен: если в кластерном анализе объекты классифицируются на основе их различий без какой-либо предварительной информации о количестве и составе классов, то в дискриминантном анализе изначально заданы количество и состав классов, а основная задача заключается в определении того, насколько точно можно предсказать принадлежность объектов к классам при помощи данного набора дискриминантных переменных (предикторов).

С вычислительной точки зрения дискриминантный анализ очень похож на дисперсионный анализ, который сравнивает размеры вариации (изменчивости, неоднородности), обусловленной разными факторами и используется для изучения различий средних значений количественной зависимой переменной, вызванных влиянием качественных независимых переменных (факторов). Важной проблемой дискриминантного анализа является определение дискриминантных

переменных (переменных, входящих в дискриминантную функцию). Возможны два подхода. Первый предполагает одновременное введение всех переменных, в этом случае учитывается каждая независимая переменная, при этом ее дискриминирующая сила не принимается во внимание. Альтернативой является пошаговый дискриминантный анализ, при котором переменные вводятся последовательно, исходя из их способности различить (дискриминировать) группы. При пошаговом анализе «с включением» на каждом шаге просматриваются все переменные, и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная должна быть включена в модель на данном шаге, и происходит переход к следующему шагу.

При пошаговом анализе «с исключением» движутся в обратном направлении, в этом случае все переменные сначала будут включены в модель, а затем на каждом шаге будут устраняться переменные, вносящие малый вклад в различение. Тогда в качестве результата успешного анализа можно сохранить только «важные» переменные в модели, т. е. те переменные, чей вклад в дискриминацию больше остальных. Пошаговый дискриминантный анализ основан на использовании уровня значимости *F*-статистики.

Как уже отмечалось, проверка качества дискриминации (различия) основана на сравнении средних дискриминантной функции для исследуемых групп. Эти средние играют настолько важную роль в дискриминантном анализе, что получили свое название — центроиды. Центроидов столько, сколько групп, т. е. один центроид для каждой группы. Кроме этого, значения дискриминантной функции также имеют свое название — дискриминантные показатели.

Прежде чем интерпретировать дискриминантную функцию следует убедиться в ее статистической значимости. Для этого проверяют нулевую гипотезу о равенстве центроидов во всех группах (чтобы дискриминантная функция была статистически значимой, эта гипотеза должна быть отвергнута). Эта гипотеза проверяется с помощью коэффициента Уилкса.

Аналогично регрессионному анализу для оценки относительной важности переменных в установлении различий между группами используются стандартизированные (нормированные) коэффициенты дискриминантной функции. В дискриминантном анализе для решения этой задачи также используются разности средних значений каждой переменной в группах и показатели дискриминантной нагрузки,

которые часто называют структурными коэффициентами корреляции — это парные линейные коэффициенты корреляции между каждой переменной U и дискриминантной функцией (z_1, z_2, …, z_k).

Важно отметить, что для получения правильных, адекватных действительности результатов, необходимо выполнение одной из предпосылок дискриминантного анализа (так же как и в регрессионном анализе), а именно отсутствие мультиколлинеарности, т. е. отсутствие связи (слабая корреляция) между переменными z_1, z_2, …, z_k. При наличии мультколлинеарности между предсказывающими переменными не существует однозначной меры относительной важности переменных.

Следует отметить, что в качестве дискриминантных переменных могут выступать не только исходные (наблюдаемые) признаки, но и главные компоненты или главные факторы (в этом случае дискриминантному анализу предшествует факторный анализ, речь о котором пойдет в следующей главе, и который позволяет сократить массив переменных и выделить новые некоррелируемые факторы).

Кроме предположения о мультиколлинеарности для корректного применения дискриминантного анализа также должны выполняться предпосылки нормальности распределения независимых переменных и однородности дисперсий (проверяется с помощью M-статистики Бокса).

Когда определен окончательный вид дискриминантной функции, можно переходить к решению задачи классификации. Сразу следует отметить, что для корректного применения дискриминантной функции для решения этой задачи должны использоваться две выборки: одна для вычисления дискриминантной функции (ее называют анализируемой), вторая — проверочная, которую используют для проверки результатов расчета на основании первой выборки. Такую процедуру проверки называют кросс-проверкой (перекрестной проверкой).

Для нового объекта находится его проекция на дискриминантную ось (т. е. значение дискриминантной функции — дискриминантный показатель) и определяется, к какому из центроидов (для первой или второй группы) он более близко расположен. В дальнейшем, он и будет отнесен к этой группе. Степень «близости» может определяться с помощью пороговых значений, которые могут определяться по следующему правилу. Если размеры групп равны, то пороговое значение

— это среднее арифметическое двух центроидов, если же группы не равны, то пороговым значением является средневзвешенная.

Также могут быть вычислены вероятности противоположных событий «объект принадлежит группе 1» и «объект принадлежит группе 2», которые в сумме дают 1 (100%). Затем на основании полученных вероятностей происходит классификация объектов.

Качество классификации оценивается с помощью так называемой «классификационной матрицы», которую еще называют смешанной матрицей или матрицей предсказания. Эта матрица содержит ряд правильно и ошибочно классифицированных случаев. Доля общего количества правильно классифицированных случаев называется коэффициентом результативности. Этот коэффициент варьируется в пределах от 50% до 100%. Нижний предел определяется тем, что даже при случайном отнесении некоторого наблюдения к одной из двух имеющихся групп (например, с помощью бросания монеты), корректность классификации составила бы 50%. Поэтому на практике удовлетворительной считается классификация с коэффициентом результативности не меньше 70%.

2.5. Основные этапы дискриминантного анализа для двух групп

Рассмотрим первый вид дискриминантного анализа, т. е. случай, когда совокупность объектов разбита на две группы.

Пусть имеется некоторая совокупность, состоящая из n объектов: Y_1, Y_2, …, Y_n. Каждый объект характеризуется m количественными признаками: P_1, P_2, …, P_m. Предположим, что объекты данной совокупности разделены, например, с помощью кластерного анализа, на две однородные группы: T и H, при этом число объектов в группе T равно n_1, в группе H – n_2. Очевидно, что $n_1 + n_2 = n$.

Объекты Y_i, попавшие в группу T обозначим через T_1, T_2, …, T_{n1}; а объекты Y_i, попавшие в группу H обозначим через H_1, H_2, …, H_{n2}.

Все значения признаков, т. е. матрицу

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

разделим на две матрицы T и H, размерностей n_1xm и n_2xm соответственно. Таким образом, каждая матрица будет состоять из значений признаков объектов, попавших в соответствующую группу.

Предположим, что после разделения на группы появился новый объект Y_q со своими значения признаков P_1, P_2, …, P_m: x_{q1}, x_{q2}, …, x_{qm}. Необходимо выяснить, к какой группе он может быть отнесен.Такая задача является основной задачей дискриминантного анализа для двух групп. Решение задачи состоит из ряда этапов.

На первом этапе строится дискриминантная функция, используя значения признаков объектов, представленных в двух образованных группах.

На втором этапе находят значение дискриминантной константы, которая определяет «разделительную линию» между двумя группами.

Третий этап связан непосредственно с определением нового объекта в конкретную группу.

Перейдем к подробному рассмотрению описанных этапов.

Дискриминантная функция, которая определяется на первом этапе, как было отмечено ранее, имеет вид

$$U = a_1z_1 + a_2z_2 + \ldots + a_mz_m.$$

Задача первого этапа состоит в нахождении неизвестных значений коэффициентов, используя значения признаков объектов. Для ее решения используют элементы матричного исчисления, а сам процесс решения состоит из следующих шагов.

1.1. По каждой группе для каждого признака находим, так называемые векторы средних, которые записываем в виде матрицы-столбца:

$$\bar{T} = \begin{pmatrix} \bar{t}_1 \\ \bar{t}_2 \\ \cdots \\ \bar{t}_m \end{pmatrix}; \qquad \bar{H} = \begin{pmatrix} \bar{h}_1 \\ \bar{h}_2 \\ \cdots \\ \bar{h}_m \end{pmatrix},$$

где значение элемента матрицы есть среднее арифметическое значений каждого признака объектов соответствующей группы.

1.2. Находим вектор разностей средних:

$$\bar{T} - \bar{H} = \begin{pmatrix} \bar{t}_1 - \bar{h}_1 \\ \bar{t}_2 - \bar{h}_2 \\ \cdots \\ \bar{t}_m - \bar{h}_m \end{pmatrix}.$$

1.3. Для каждой группы находим ковариационную матрицу:

$$S_t = \begin{pmatrix} s_{11}^t & s_{12}^t & \dots & s_{1m}^t \\ s_{21}^t & s_{22}^t & \dots & s_{2m}^t \\ \dots & \dots & \dots & \dots \\ s_{m1}^t & s_{m2}^t & \dots & s_{mm}^t \end{pmatrix}; \quad S_h = \begin{pmatrix} s_{11}^h & s_{12}^h & \dots & s_{1m}^t \\ s_{21}^h & s_{22}^h & \dots & s_{2m}^h \\ \dots & \dots & \dots & \dots \\ s_{m1}^h & s_{m2}^h & \dots & s_{mm}^h \end{pmatrix}.$$

Обе матрицы являются симметричными относительно главной диагонали. Каждый элемент определяет значение коэффициента ковариации между соответствующими признаками.

1.4. Находим суммарную ковариационную матрицу по формуле:

$$S = \frac{1}{n_1 + n_2 - 2}(n_1 \cdot S_t + n_2 \cdot S_h).$$

Матрица S – квадратная, порядка m.

1.5. Находим матрицу S^{-1}, обратную к матрице S.

1.6. Находим коэффициенты дискриминантной функции в виде матрицы-столбца (вектор коэффициентов):

$$A = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_m \end{pmatrix} = S^{-1} \cdot (\overline{T} - \overline{H}).$$

После нахождения неизвестных коэффициентов дискриминантной функции переходим ко второму этапу.

Процесс нахождения дискриминантной константы состоит из следующих шагов.

2.1. Находим оценки дискриминантной функции для двух кластер-групп:

$$U_t = T \cdot A = \begin{pmatrix} u_1^t \\ u_2^t \\ \dots \\ u_{n_1}^t \end{pmatrix}; \qquad U_h = H \cdot A = \begin{pmatrix} u_1^h \\ u_2^h \\ \dots \\ u_{n_2}^h \end{pmatrix}.$$

2.2. Находим средние значения оценок $\overline{U}_t$; $\overline{U}_h$.

2.3. Находим значение константы дискриминации как среднее значение найденных средних оценок:

$$C = \frac{1}{2}(\overline{U}_t + \overline{U}_h).$$

После нахождения константы дискриминации переходим к третьему этапу, а именно — определяем группу, к которой отнесем новый объект. Третий этап состоит из двух шагов.

3.1. Находим значения дискриминантной функции для нового объекта:

$$U(Y_q) = a_1 x_{q1} + a_2 x_{q2} + \ldots + a_m x_{qm}.$$

3.2. Определяем принадлежность нового объекта к одной из групп сравнением полученного значения дискриминантной функции с константой дискриминации по следующему правилу: если $U(Y_q) > C$, то новый объект попадает в группу, для которой средняя оценок (см. п. 2.2) больше. В противном случае — наоборот.

Рассмотрим пример применения дискриминантного анализа.

Предположим, что оценивается деятельность девяти промышленных предприятий по двум показателям (признакам):

P_1 — среднесписочная численность персонала (в тыс. чел.);

P_2 — балансовая прибыль (в *млн* руб.).

В соответствии со значениями этих показателей все девять предприятий были поделены, например, с помощью кластерного анализа, на две группы: T — группа, состоящая из предприятий, которые оцениваются как стабильные, успешные и преуспевающие; H — группа, состоящая из остальных, менее успешных предприятий.

Результаты распределения предприятий по группам представлены в таблице.

Таблица 2.2

Группы предприятий в зависимости от значений признаков

Группы предприятий	*Объекты*	*Численность ППП P_1*	*Балансовая прибыль P_2*
Успешные предприятия (T)	T_1	14,115	22,981
	T_2	14,904	21,481
	T_3	13,627	28,669
	T_4	10,545	10,199
Остальные, менее успешные предприятия (H)	H_1	4,428	11,124
	H_2	5,510	6,091
	H_3	4,214	11,842
	H_4	5,527	11,873
	H_5	4,211	12,860

Требуется решить, можно ли к группе успешных отнести предприятие, имеющее следующие показатели: численность – 9,592 тыс. чел. и прибыль – 12,840 *млн* руб.

Для решения этой задачи проведем дискриминантный анализ в соответствии с описанными выше этапами.

Первый этап. Построение дискриминантной функции:

1.1. Находим векторы средних: $\bar{T} = \begin{pmatrix} 13{,}29775 \\ 20{,}8325 \end{pmatrix}$; $\bar{H} = \begin{pmatrix} 4{,}778 \\ 10{,}758 \end{pmatrix}$.

1.2. Находим вектор разностей средних: $\bar{T} - \bar{H} = \begin{pmatrix} 4{,}778 \\ 10{,}758 \end{pmatrix}$.

1.3. Находим для каждой группы ковариационную матрицу:

$S_t = \begin{pmatrix} 2{,}7335 & 8{,}6623 \\ 8{,}6623 & 44{,}8797 \end{pmatrix}$ $S_h = \begin{pmatrix} 0{,}3718 & -0{,}9025 \\ -0{,}9025 & 5{,}7503 \end{pmatrix}$.

1.4. Находим суммарную ковариационную матрицу:

$$S = \begin{pmatrix} 1{,}8276 & 4{,}3052 \\ 4{,}3052 & 29{,}7529 \end{pmatrix}.$$

1.5. Находим обратную матрицу к суммарной ковариационной:

$$S^{-1} = \begin{pmatrix} 0{,}8302 & -0{,}1201 \\ -0{,}1201 & 0{,}051 \end{pmatrix}.$$

1.6. Находим вектор коэффициентов дискриминантной функции:

$$\mathrm{A} = S^{-1} * (\bar{T} - \bar{H}) = \begin{pmatrix} 5{,}8626 \\ -0{,}5097 \end{pmatrix}.$$

Таким образом, дискриминантная функция имеет вид:

$$U(Y_k) = a_1 x_{1k} + a_2 x_{2k} = 5{,}8626 * x_{1k} + (-0{,}5097) * x_{2k}.$$

Второй этап. Нахождение дискриминантной константы:

2.1. Находим оценки дискриминантной функции для двух кластер-групп:

$$U_t = T \cdot A = \begin{pmatrix} 71{,}0365 \\ 76{,}4266 \\ 65{,}2764 \\ 56{,}6222 \end{pmatrix} \qquad U_h = H \cdot A = \begin{pmatrix} 20{,}2895 \\ 29{,}1981 \\ 18{,}6689 \\ 26{,}3506 \\ 18{,}1324 \end{pmatrix}.$$

2.2. Находим средние оценок: $\bar{U}_t = 67{,}3404$ $\bar{U}_h = 22{,}5279$.

2.3. Находим значение константы дискриминации:

$$\mathrm{C} = 44{,}9342.$$

Третий этап. Определение объекта в имеющиеся группы:

3.1. Находим значение дискриминантной функции для значений показателей нового объекта:

$$U(Y_q) = a_1 x_{1q} + a_2 x_{2q} = 5{,}8626 * 9{,}592 + (-0{,}5097) * 12{,}840 = 49{,}689$$

3.2. Определяем группу сравнением значения дискриминантной функции со значением константы дискриминации:

$$U(Y_k) = 49{,}689 > \mathrm{C} = 44{,}9342,$$

следовательно, объект должен быть отнесен к первой группе, т. е. группе успешных предприятий, так как выполнено неравенство $\overline{U}_t = 67{,}3404 > \overline{U}_h = 22{,}5279$.

Кроме рассмотренной выше задачи, дискриминантный анализ позволяет проверить правильность первоначального распределения на группы. Поясним последнее на предыдущем примере.

Предположим, что принадлежность объекта T_4 к успешным предприятиям у исследователя вызывает некоторые сомнения. Проверку этого факта можно провести следующим образом:

– убираем объект T_4 из рассмотрения;

– строим дискриминантную функцию для оставшихся восьми предприятий;

– находим константу дискриминации;

– находим значение новой дискриминантной функции для показателей объекта T_4;

– выясняем принадлежность объекта T_4 к одной из групп, сравнением соответствующих величин.

Таким образом, рассмотренные кластерный и дискриминантный анализ работают, как правило, с объектами (людьми, банками, хозяйствующими субъектами, странами и т. д.), разделяя их на однородные группы. Работой с признаками (ВВП, прибыль и т. д.), характеризующими объекты, занимается факторный и компонентный анализ.

3. ФАКТОРНЫЙ И КОМПОНЕНТНЫЙ АНАЛИЗ

Факторный анализ (*factor analysis*) и компонентный анализ (*components analysis*) — это многомерные статистические методы снижения размерности, применяемые для изучения взаимосвязей между значениями количественных переменных. Основная идея факторного анализа заключается в том, что имеющиеся зависимости между большим числом исходных наблюдаемых переменных определяются существованием гораздо меньшего числа скрытых или латентных переменных, называемых факторами. Задача компонентного анализа состоит в преобразовании исходной системы взаимосвязанных переменных в новую систему некоррелированных обобщенных показателей или ортогональных показателей. Новые некоррелированные показатели называются компонентами.

3.1. Сравнение основных понятий и целей факторного и компонентного анализа

Очень часто под термином «факторный анализ» понимают методы факторного и компонентного анализов. На самом деле, факторный и компонентный анализы — методы снижения размерности. Оба метода решают в принципе одну и ту же задачу и поэтому результаты похожи. На первом этапе применения этих методов строится так называемая «матрица нагрузок», в которой первый общий фактор и первая главная компонента, как правило, совпадают. Второй фактор может уже существенно отличаться от второй главной компоненты и т. д. Иногда даже число общих факторов может существенно отличаться от числа весомых главных компонент.

Предпосылки методов факторного и компонентного анализа различаются.

Цель компонентного анализа — объяснить всю корреляцию между переменными и всю суммарную дисперсию исходных переменных. Число первоначально извлеченных компонент совпадает с числом исходных переменных. Хотя в дальнейшем анализе используются только главные компоненты.

В факторном анализе с самого начала предполагается, что число извлеченных факторов будет существенно меньше числа первоначальных переменных. В факторном анализе извлеченные новые пе-

ременные-факторы в принципе не могут объяснить полностью суммарную дисперсию исходных переменных и их корреляции.

Компонентный анализ используют чаще как метод избавления от мультиколлинеарности объясняющих переменных в регрессионном анализе. Главные компоненты всегда не коррелированы между собой.

Общие факторы в факторном анализе могут быть коррелированными или слабо коррелированными между собой.

В факторном анализе делается больший акцент на интерпретации факторов, а в компонентном — на сокращение размерности пространства за счет некоррелированных переменных.

Факторный анализ вначале был применен в психологии, когда возникла необходимость свести при обработке большое число психологических тестов к небольшому числу факторов, объясняющих способности интеллекта. Основная идея факторного анализа была сформулирована еще Ф. Гальтоном, основоположником измерений индивидуальных различий. Она сводится к тому, что если несколько признаков, измеренных на группе индивидов, изменяются согласованно, то можно предположить существование одной общей причины этой совместной изменчивости — фактора как скрытой (латентной), непосредственно недоступной измерению переменной. В настоящее время факторный анализ широко используется для решения практических задач не только в психологии, но и в социологии, политологии, экономике, медицине, маркетинге, химии и других областях, а также при решении проблем информационной и экономической безопасности.

В дальнейшем под термином «факторный анализ» будем понимать некое соединение метода факторного и компонентного анализа.

Как общенаучный метод факторный (компонентный) анализ становится средством для замены набора коррелирующих измерений существенно меньшим числом новых переменных-факторов (компонент). При этом основными требованиями являются, во-первых, минимальная потеря информации, содержащейся в исходных данных, во-вторых, возможность представления (интерпретации) факторов через исходные переменные.

Таким образом, главная цель факторного анализа — уменьшение размерности исходных данных с целью их экономного описания при условии минимальных потерь исходной информации. Результатом факторного анализа является переход от множества исходных переменных к существенно меньшему числу новых переменных —

факторов (главных компонент). Фактор при этом интерпретируется как причина совместной изменчивости нескольких исходных переменных.

С учетом вышеизложенного, методами факторного анализа могут быть решены следующие задачи:

1) отыскание скрытых, но объективно существующих закономерностей исследуемого процесса, определяемых воздействием внутренних и внешних причин;

2) описание изучаемого процесса значительно меньшим числом факторов по сравнению с первоначально взятым количеством признаков;

3) выявление первоначальных признаков, наиболее тесно связанных с основными факторами;

4) прогнозирование процесса на основе уравнения регрессии, построенного по полученным факторам.

Факторный анализ дает возможность количественно определить нечто непосредственно не измеряемое, исходя из нескольких доступных измерению переменных.

Например, характеристики «посещает развлекательные мероприятия», «много разговаривает», «охотно идет на контакт с любым незнакомым человеком» могут служить оценками качества «общительность», которое непосредственно не поддается количественному измерению.

В начале применения факторного и компонентного анализа используется алгоритм метода главных компонент.

3.2. Основные понятия метода главных компонент

Метод главных компонент состоит в последовательном извлечении компонент. Для единственности решения в этом методе компоненты должны быть упорядочены по убыванию доли объясняемой суммарной дисперсии исходных переменных. Первая компонента характеризует наибольшую долю вариации исходных переменных, вторая компонента объясняет наибольшую долю дисперсии, не объясняемой первой компонентой и т. д. В результате компонентного анализа число полученных некоррелированных компонент совпадает с числом исходных переменных, т. е. классический компонентный анализ сохраняет размерность пространства переменных. Однако затем, сре-

ди всех компонент выбирают главные, и дальнейшее исследование производится уже на главных компонентах.

Каждой извлеченной компоненте соответствует характеристика, называемая собственным значением. Собственное значение показывает часть вариации исходных переменных, объясняемую компонентой. В компонентном анализе, если используется корреляционная матрица, каждая переменная стандартизирована и ее дисперсия равна 1. Следовательно, если число исходных переменных k, то суммарная дисперсия равна k. Компонентный анализ сохраняет всю суммарную дисперсию, поэтому сумма всех собственных значений равна числу исходных переменных.

На основе полученных собственных значений рассчитывается матрица нагрузок и дается интерпретация компонент. На основе полученной матрицы нагрузок в компонентном анализе может быть произведено вращение факторов для получения простой структуры матрицы нагрузок. Классический компонентный анализ, в отличие от факторного анализа, не предполагает вращение матрицы нагрузок. Но вращение настолько полезная возможность методов снижения размерности, что алгоритмы вращения были разработаны и для компонентного анализа.

На основе окончательной матрицы нагрузок рассчитывают индивидуальные значения главных компонент для каждого объекта наблюдения. Индивидуальные значения главных компонент на объектах представляют собой линейную комбинацию исходных переменных для каждого фактора. На практике для интерпретации и дальнейшего анализа используют компоненты, удовлетворяющие следующим условиям:

– их собственные значения должны быть больше 1 – это означает, что компонента более информативна, чем стандартизированная переменная;

– компонента должна иметь хотя бы одну нагрузку больше критического значения — это означает, что компонента тесно связана, по крайней мере, с одной исходной переменной.

Отобранные для дальнейшего анализа компоненты называют главными компонентами. Отсюда название метода — метод главных компонент.

Индивидуальные значения главных компонент могут быть использованы для дальнейшего статистического анализа, например, для

построения уравнения регрессия на главных компонентах или классификации наблюдений по главным компонентам.

Основные этапы метода главных компонент

Задача метода главных компонент состоит в определении того, насколько необходимо выделить компонент и какие они должны быть, чтобы наиболее точно с их помощью можно было воспроизвести и объяснить исследуемые связи, которые представлены в виде корреляционной матрицы. Метод главных компонент основан на матричных преобразованиях. Исходными элементами для метода главных компонент являются объекты и признаки, их характеризующие.

Итак, имеется некоторая совокупность, состоящая из n объектов: Y_1, Y_2, …, Y_n. Каждый объект характеризуется m количественными признаками: P_1, P_2, …, P_m.

Все значения признаков, как и ранее для исходных данных кластерного анализа, представим в виде следующей матрицы:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix},$$

где x_{ij} – значение признака P_j для объекта Y_i.

Таким образом, в матрице X, размерности nxm, строки соответствуют объектам, столбцы – признакам.

На первом этапе для каждого признака P_j находим средние значения $\bar{x}_j$ и среднеквадратические отклонения s_j по формулам:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}; \qquad s_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n}}.$$

Затем строим матрицу нормированных значений Z:

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{pmatrix},$$

где $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$.

На втором этапе происходит формирование матрицы парных коэффициентов корреляции R, размерности mxm:

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ r_{21} & r_{22} & \dots & r_{2m} \\ \dots & \dots & \dots & \dots \\ r_{m1} & r_{m2} & \dots & r_{mm} \end{pmatrix},$$

где каждый элемент матрицы показывает силу взаимосвязи между двумя соответствующими признаками и определяется по формуле:

$$r_{jl} = \frac{\overline{x_j \cdot x_l} - \bar{x}_j \cdot \bar{x}_l}{s_j \cdot s_l}; \quad \overline{x_j \cdot x_l} = \frac{\sum_{i=1}^{n} x_{ij} x_{il}}{n}.$$

Необходимо отметить, что элементы матрицы R быстро и достаточно просто можно найти с помощью стандартных процедур Excel.

На третьем этапе для матрицы R находим собственные значения и соответствующие этим значениям собственные векторы. Для нахождения собственных чисел λ необходимо решить уравнение:

$$|\lambda \cdot E - R| = 0,$$

где E — единичная матрица, порядка m.

Решение такого уравнения может представлять определенные трудности, поэтому собственные числа и собственные векторы могут быть найдены с использованием онлайн-калькулятора, например, https://matrixcalc.org/vectors.html

Полученные собственные числа расположим в порядке убывания:

$$\lambda_1 > \lambda_2 > \dots > \lambda_m.$$

Из собственных чисел составим матрицу собственных значений:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_m \end{pmatrix}.$$

Соответствующие собственные векторы обозначим через V_1, V_2, …, V_m.

В дальнейшем, будут использоваться нормированные собственные векторы, которые достаточно просто получить из данных V_1, V_2,

…, V_m, поделив каждую координату вектора на длину этого вектора. Произведя эту операцию с каждым из собственных векторов, соединим их в матрицу L. Таким образом, матрица L — это матрица размерности mxm, столбцы которой являются значениями нормированных собственных векторов.

На четвертом этапе находим вклад каждой главной компоненты в суммарную дисперсию исходных признаков K, которая равна сумме всех собственных чисел:

$$K = \sum_{i=1}^{m} \lambda_i.$$

Вклад каждой компоненты определяется по формуле:

$$\frac{\lambda_i}{K} \cdot 100\%.$$

Необходимо отметить, что наибольший вклад в общую дисперсию вносит первая компонента, а наименьший — последняя.

На этом этапе возможно выделение главных компонент, которые оказывают наибольшее влияние на общую вариацию. Для определения оптимального числа главных компонент существует два основных критерия. По критерию Кайзера значимыми считаются компоненты (факторы), для которых значения собственных чисел превышают 1. По критерию Кеттеля значимыми считаются факторы, которые объясняют более 90% всей дисперсии.

На пятом этапе построим матрицу факторных нагрузок:

$$A = L \cdot \Lambda^{\frac{1}{2}} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix},$$

где матрица $\Lambda^{\frac{1}{2}}$ — это матрица, полученная из Λ извлечением квадратного корня из диагональных элементов.

Правильность получения матрицы A можно проверить по выполнению следующих равенств:

$$AA^T = R;\ A^T A = \Lambda.$$

Элементы матрицы факторных нагрузок являются коэффициентами корреляции между главными компонентами (факторами) и исходными признаками.

На шестом этапе строим матрицу факторных весов (матрицу значений главных компонент) по формуле:

$$F = Z(A^T)^{-1}.$$

Анализируя элементы данной матрицы можно выяснить, например, какие исходные признаки могут быть соединены во вновь полученные, ранее латентные факторы.

Приведем пример использования метода главных компонент.

Пример. При переходе школьников из начальной школы в среднюю было необходимо распределить их в профильные классы или оставить в обычном классе. Для этого была разработана система тестов, которая позволила оценить способности младшего школьника. Фрагмент результатов тестов для девяти школьников приведен в таблице.

Таблица 3.1

Результаты оценок характеристик и способностей школьников

	Рост, см	*Вес, кг*	*Память на числа*	*Логическое мышление**	*Техника чтения*	*Фантазия**	*Прыжки в длину*
Иванов	150	35	7	1	180	0	70
Костин	156	34	6	0	156	1	120
Мишин	168	32	5	1	189	0	140
Денисов	160	39	2	0	190	1	80
Юрин	162	45	4	0	200	1	90
Толин	159	30	5	0	230	0	100
Кирин	145	28	9	1	250	1	120
Олегов	150	32	7	1	200	0	110
Павлов	160	25	7	1	180	0	100

* 1 – есть, 0 – нет.

Требуется определить, какой школьник лучше всего подходит для соответствующего профильного класса.

Для решения поставленной задачи используем метод главных компонент. Все расчеты будем производить средствами Excel.

Исходные данные: $n = 9$; $m = 7$ и данные таблицы 3.1.

Найдем среднее и среднеквадратическое (стандартное) отклонение для каждого признака (см. таблицу 3.2.), затем сформируем матрицу-таблицу нормированных значений признаков объектов.

Таблица 3.2

Значения средних и стандартных отклонений признаков

Признак	P_1	P_2	P_3	P_4	P_5	P_6	P_7
Среднее	156,67	33,33	5,78	0,56	197,22	0,44	103,33
Ст.от.	7,16	5,96	2,05	0,53	28,03	0,53	21,8

Таблица 3.3

Нормированные значения признаков объектов

	Рост	*Вес*	*Память*	*Логика*	*Чтение*	*Фантазия*	*Прыжки*
Иванов	-0,93	0,28	0,60	0,84	-0,61	-0,84	-1,53
Костин	-0,09	0,11	0,11	-1,05	-1,47	1,05	0,76
Мишин	1,58	-0,22	-0,38	0,84	-0,29	-0,84	1,68
Денисов	0,47	0,95	-1,84	-1,05	-0,26	1,05	-1,07
Юрин	0,74	1,96	-0,87	-1,05	0,10	1,05	-0,61
Толин	0,33	-0,56	-0,38	-1,05	1,17	-0,84	-0,15
Кирин	-1,63	-0,90	1,57	0,84	1,88	1,05	0,76
Олегов	-0,93	-0,22	0,60	0,84	0,10	-0,84	0,31
Павлов	0,47	-1,40	0,60	0,84	-0,61	-0,84	-0,15

На втором этапе находим матрицу-таблицу R, состоящую из парных коэффициентов.

Таблица 3.4

Значения парных коэффициентов корреляции между признаками

	Рост	Вес	Память	Логика	Чтение	Фантазия	Прыжки
Рост	1,00	0,26	-0,70	-0,34	-0,36	-0,12	0,18
Вес	0,26	1,00	-0,67	-0,58	-0,23	0,50	-0,42
Память	-0,70	-0,67	1,00	0,71	0,27	-0,24	0,30
Логика	-0,34	-0,58	0,71	1,00	0,11	-0,55	0,25
Чтение	-0,36	-0,23	0,27	0,11	1,00	0,06	0,12
Фантазия	-0,12	0,50	-0,24	-0,55	0,06	1,00	-0,04
Прыжки	0,18	-0,42	0,30	0,25	0,12	-0,04	1,00

Необходимо отметить наличие мультиколлинеарности между признаками, что может привести при применении методов корреляционно-регрессионного анализа к искаженным результатам.

На третьем этапе с помощью онлайн-калькулятора находим собственные числа и собственные векторы матрицы R.

Перенумеруем собственные числа в порядке убывания их значений:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
3,065	1,483	1,075	0,737	0,377	0,176	0,087

и сформируем матрицу-таблицу Λ собственных чисел:

Таблица 3.5

Матрица собственных значений

$\Lambda =$

	1	2	3	4	5	6	7
1	3,065	0	0	0	0	0	0
2	0	1,483	0	0	0	0	0
3	0	0	1,075	0	0	0	0
4	0	0	0	0,737	0	0	0
5	0	0	0	0	0,377	0	0
6	0	0	0	0	0	0,176	0
7	0	0	0	0	0	0	0,087

Для каждого собственного числа был найден собственный вектор. Перейдем к рассмотрению собственных векторов.

Таблица 3.6

Собственные векторы

V_1	V_2	V_3	V_4	V_5	V_6	V_7
-1,428	2,072	0,303	-0,791	0,655	-0,991	-2,301
-2,17	-0,639	-0,149	0,309	4,457	1,237	-0,667
2,35	-0,637	-0,034	1,14	-0,12	0,524	-3,191
2,166	0,544	-0,278	0,358	4,137	-1,329	0,851
0,934	-1,439	0,484	-2,966	1,142	-0,113	-0,353
-1,285	-1,74	0,516	1,667	0,109	-1,574	-0,196
1	1	1	1	1	1	1

После определения модуля (длины вектора) и деления каждого элемента вектора на его длину, получаем следующую матрицу-таблицу 3.7. нормированных собственных векторов:

Таблица 3.7

Нормированные собственные векторы

	V_1	V_2	V_3	V_4	V_5	V_6	V_7
	-0,316	0,611	0,233	-0,206	0,104	-0,349	-0,545
	-0,479	-0,188	-0,115	0,081	0,707	0,436	-0,158
$L =$	0,519	-0,188	-0,026	0,297	-0,019	0,185	-0,756
	0,479	0,160	-0,214	0,093	0,656	-0,469	0,202
	0,206	-0,424	0,372	-0,773	0,181	-0,040	-0,084
	-0,284	-0,513	0,397	0,434	0,017	-0,555	-0,046
	0,221	0,295	0,769	0,261	0,159	0,353	0,237

На четвертом этапе находим вклады компонент в общую дисперсию, формируем полученные данные в виде таблицы 3.8.

Таблица 3.8

Вклады компонент в общую дисперсию

	Собственные числа	*% объясненной дисперсии*	*Накопленный %*
λ_1	3,065	43,79	43,79
λ_2	1,483	21,19	64,98
λ_3	1,075	15,36	80,34
λ_4	0,737	10,53	90,87
λ_5	0,377	5,39	96,26
λ_6	0,176	2,51	98,77
λ_7	0,087	1,23	100,00

По критерию Кайзера значимыми являются три фактора, у которых значения собственных чисел превышает 1. Суммарно они объясняют немного более 80% общей дисперсии.

По критерию Кеттеля можно считать значимыми четыре фактора, которые объясняют более 90% общей дисперсии.

Отсюда можно сделать вывод, что для дальнейшего исследования различных взаимосвязей достаточно рассматривать либо 3, либо 4 фактора вместо первоначальных 7.

На пятом этапе после умножения соответствующих матриц, получаем матрицу факторных нагрузок.

Таблица 3.9

Матрица факторных нагрузок

	$Ф_1$	$Ф_2$	$Ф_3$	$Ф_4$	$Ф_5$	$Ф_6$	$Ф_7$
Рост	-0,55	0,74	-0,24	-0,18	0,06	-0,15	-0,16
Вес	-0,44	0,12	0,23	0,07	0,23	0,18	-0,05
Память	0,91	0,03	0,23	0,26	-0,01	0,08	-0,22
Логика	0,84	0,22	-0,20	0,08	0,10	-0,20	0,06
Чтение	0,36	-0,38	0,52	-0,26	0,11	-0,02	-0,02
Фантазия	-0,50	-0,41	0,63	0,37	0,01	-0,23	-0,01
Прыжки	0,39	-0,80	-0,36	0,22	0,10	0,15	0,07

Анализ таблицы 3.9. показывает, что первый фактор сильнее всего связан с признаками, характеризующими память и логическое мышление, т. е. он может отвечать за способности школьников к техническим дисциплинам.

Второй фактор теснее всего связан с признаками, характеризующими спортивное развитие школьника (рост и прыжки).

И, наконец, третий фактор можно охарактеризовать, как фактор гуманитарной направленности (чтение и фантазия).

Связь остальных факторов с первоначальными признаками слабая, поэтому их можно не принимать во внимание.

Итак, вместо семи признаков для решения нашей задачи достаточно рассмотреть только три фактора, при этом мультиколлинеарность у новых факторов отсутствует.

Для окончательного решения задачи необходимо перейти к последнему этапу и построить матрицу факторных весов.

Таблица 3.10

Матрица факторных весов

	Математический класс, $Ф_1$	*Спортивный класс, $Ф_2$*	*Гуманитарный класс, $Ф_3$*
Иванов	-0,74	-1,45	-0,94
Костин	1,31	1,02	-0,70
Мишин	-1,03	1,60	1,44
Денисов	0,26	-1,09	0,90
Юрин	0,27	-0,38	0,70
Толин	1,79	-0,15	0,28
Кирин	-0,80	0,55	-1,19
Олегов	-0,29	0,52	-1,15
Павлов	-0,78	-0,61	0,66

Анализ таблицы 3.10. показывает, что в математический класс можно рекомендовать Костина и Толина, в спортивный — Мишина, Кирина и Олегова, в гуманитарный — Денисова, Юрина и Павлова. Иванова лучше отправить в обычный, непрофильный класс, так как все выделенные способности у него ниже среднего.

До настоящего момента для исследования использовались данных, явно не зависящие от времени, т. е. среди признаков, характеризующих объекты, не было временного показателя. Однако существуют проблемы, в которых требуется выяснить, как меняется значение какого-либо показателя в определенные моменты времени. Для решения таких задач применяются временные ряды.

4. ВРЕМЕННЫЕ РЯДЫ И ИХ ПРИМЕНЕНИЕ

Для проведения математико-статистического исследования какой-либо проблемы, как было отмечено ранее, необходимо наличие статистических данных. Все такие данные можно разделить на два вида. Данные первого вида характеризуют множество различных объектов и их признаков в определенный момент времени. Данные второго вида характеризуют, как правило, один объект, но за несколько последовательных периодов (моментов) времени.

Математико-статистическое исследование с использованием данных первого типа проводится методами корреляционно-регрессионного анализа, рассмотренными во второй части методических рекомендаций.

Второй тип данных применяется при исследовании особых моделей — моделей временных рядов, с помощью которых проводят статистическое описание процессов во времени.

Цель анализа временных рядов — понять и описать механизм, который порождает значения ряда, а после этого попробовать предсказать поведение ряда, по крайней мере, в ближайшем будущем.

4.1. Понятие и основные элементы временного ряда

Временной ряд (ряд динамики) — это последовательность значений какого-либо показателя, упорядоченная в хронологическом порядке, т. е. в порядке возрастания временного параметра.

Временные ряды состоят из двух элементов:

– периода времени, за который или по состоянию на который фиксируются числовые значения;

– числовых значений того или иного показателя, называемых уровнями ряда.

Временные ряды имеют два главных отличия от пространственных выборок, рассмотренных ранее. Во-первых, уровни временного ряда, как правило, не являются статистически независимыми и, во-вторых, элементы временного ряда не являются одинаково распределенными.

Классификация временных рядов

Временные ряды классифицируются по следующим признакам:

1) по форме представления уровней существуют:

– ряды абсолютных показателей;

– относительных показателей;

– средних величин.

2) по количеству показателей, для которых определяются уровни в каждый момент времени, существуют:

– одномерные временные ряды;

– многомерные временные ряды.

3) по характеру временного параметра существуют:

– моментные временные ряды;

– интервальные временные ряды.

В моментных временных рядах уровни характеризуют значения показателя по состоянию на определенные моменты времени (например, курс акций на определенные моменты времени). В интервальных рядах уровни характеризуют значение показателя за определенные периоды времени (например, прибыль предприятия за определенные месяцы).

4) по расстоянию между датами и интервалами времени выделяют:

– равноотстоящие ряды, в которых даты регистрации или окончания периодов следуют друг за другом с равными интервалами;

– не равноотстоящие ряды, в которых принцип равных интервалов не соблюдается.

5) по наличию пропущенных значений существуют:

– полные временные ряды;

– неполные временные ряды.

6) по наличию случайностей различного рода бывают:

– детерминированные ряды, которые получают на основе значений некоторой неслучайной функции (ряд последовательных данных о количестве дней в месяцах);

– случайные ряды, которые можно рассматривать как результат реализации некоторой случайной величины.

Очевидно, что анализу, а затем и прогнозированию подвергаются случайные ряды (которые и будут рассматриваться в данном разделе).

7) в зависимости от вероятностных свойств выделяют:

– стационарные ряды, в которых вероятностные свойства уровней ряда не изменяются во времени, т. е. их закон распределения и числовые характеристики постоянны и не зависят от времени;

– нестационарные ряды, в которых прослеживается зависимость вероятностных характеристик от времени.

Самым простейшим примером стационарного временного ряда является, так называемый «белый шум» — это ряд с некоррелированными ошибками и математическим ожиданием, равным нулю.

Основные составляющие уровня ряда

Формирование каждого уровня временного ряда происходит в результате воздействия множества факторов, которые условно могут быть разделены на три группы. К первой группе относятся факторы, которые формируют некоторую тенденцию ряда. Циклические или сезонные колебания формируют факторы второй группы. И, наконец, к третьей группе относятся факторы, формирующие некоторую случайную составляющую.

В связи с этим значения уровней ряда могут содержать следующие компоненты: трендовая компонента (тренд), циклическая (или сезонная) компонента, случайная составляющая.

Рассмотрим более подробно каждую из компонент уровня временного ряда.

Воздействие по отдельности факторов первой группы на исследуемый показатель, как правило, является разнонаправленным. Однако их совместное действие может привести к формированию некоторой длительной тенденции изменения показателя. Тенденция, которая характеризует совместное долговременное воздействие совокупности факторов на изменение (динамику) исследуемого показателя, является основной составляющей уровня временного ряда и называется трендом u_t.

Гипотетический временной ряд, содержащий только возрастающую тенденцию, показан на рисунке.

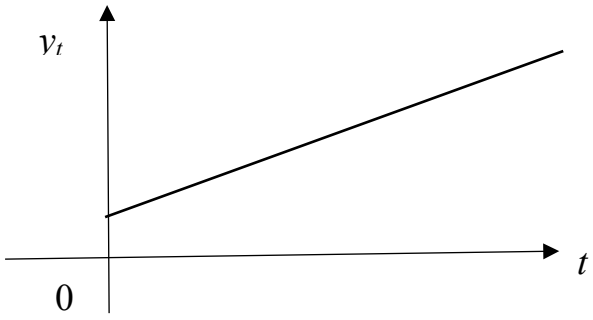

Рис. 4.1.

Влияние факторов второй группы может привести к тому, что во временных рядах могут появиться некоторые колебания, более или менее регулярные. Эти колебания, в зависимости от длительности периода колебаний, определяют циклическую или сезонную составляющую ряда.

Для циклической компоненты c_t повторяемость рассматриваемого процесса, т. е. период колебаний является достаточно длительным. Примером таких колебаний могут служить циклы Кондратьева в экономике.

Если период повторяемости (колебания) не превышает одного года, то циклическую составляющую называют сезонной v_t. Примером сезонных колебаний могут быть цены на сельскохозяйственную продукцию, так как в летний период они ниже, чем в зимний. Причины возникновения сезонных колебаний могут иметь и социальный характер (например, увеличение закупок в предпраздничные дни).

Гипотетический временной ряд, содержащий только сезонную составляющую, представлен на рисунке.

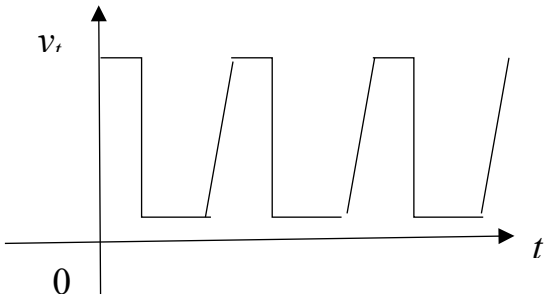

Рис. 4.2

Воздействие различных случайностей, т. е. факторов третьей группы приводит к формированию еще одной компоненты уровня ряда — случайной составляющей ε_t. Влияние этой составляющей не поддается никакому учету и регистрации.

Пример ряда, содержащего только случайную компоненту, приведен на рисунке.

Рис. 4.3

Нетрудно понять, что в реальности, рассмотренные выше примеры, встречаются крайне редко. Достаточно часто подлежат исследованию ряды, в которых уровни содержат все компоненты, при этом уровень ряда может быть представлен в виде суммы или в виде произведения компонент. В первом случае временной ряд определяет аддитивную модель, во втором — мультипликативную.

Для достижения цели анализа отдельного временного ряда необходимо решить основную задачу исследования, а именно выявить и проанализировать каждую из рассмотренных выше компонент. Полученные результаты в дальнейшем используются для прогнозирования будущих значений ряда, а также при исследовании взаимосвязи двух или более временных рядов.

4.2. Основные этапы анализа временного ряда

Основная задача исследования отдельного временного ряда формирует следующие основные этапы его анализа.

На первом, предварительном этапе исследования происходит аналитическое и, при необходимости, графическое описание временного ряда, а также выявление и, по возможности, исключение аномальных уровней.

Выделение, анализ и удаление неслучайных компонент временного ряда — тренда, циклических и сезонных составляющих — составляет содержание второго этапа исследования.

На третьем этапе исследования происходит анализ случайной составляющей.

Все результаты предыдущих этапов исследования используются на четвертом этапе, основной задачей которого является прогнозирование развития исследуемого процесса.

При наличии нескольких временных рядов, исследование взаимосвязи между ними составляет основное содержание пятого этапа анализа.

Рассмотрим первый, предварительный этап исследования.

Для получения адекватных результатов анализа временного ряда важно правильно его представить аналитически и, возможно, графически. Удобнее всего иметь дело с равноотстоящими и сопоставимыми уровнями ряда. Для успешного изучения динамики процесса временной ряд должен иметь достаточную длину и не содержать пропущенные уровни.

Предположим, что временной ряд

t	y_t
1	y_1
2	y_2
…	…
n	y_n

удовлетворяет всем описанным выше условиям.

Следующая задача состоит в выявлении так называемых аномальных уровней ряда. Под аномальным уровнем понимается уровень ряда, который выделяется среди всех остальных своим значением, т. е. его значение не характерно для динамики изучаемого процесса, но при этом оказывает существенное влияние на значение основных характеристик временного ряда.

Появление аномальных значений может быть вызвано различными техническими ошибками при сборе, записи и передаче информации. Такие ошибки должны быть выявлены и исключены, а уровни ряда, если это возможно, заменяются их истинными значениями, либо заменяются средней арифметической двух соседних уровней ряда.

Однако аномальные значения могут отражать и реальные процессы (например, скачок курса доллара или его падение). В этом случае ошибка, в результате которой получено аномальное значение уровня, не подлежит устранению. Тогда так же, как и для технических ошибок, аномальное значение уровня заменяется расчетным, но эта замена должна учитываться в дальнейшем.

Для выявления аномальных явлений применяется метод Ирвина, суть которого состоит в следующем.

Для каждого уровня ряда, начиная со второго, вычисляется значение аномального коэффициента λ_t по формуле:

$$\lambda_t = \frac{|y_t - y_{t-1}|}{\sigma_y},$$

где $\sigma_y = \sqrt{\frac{\sum_{t=1}^{n}(y_t-\bar{y})^2}{n-1}}$; $\bar{y} = \frac{\sum_{t=1}^{n} y_t}{n}$.

Полученные расчетные значения λ_2, λ_3, …, λ_n сравниваются с табличным значением $\lambda_{кр}$, которое определяется по таблице критерия Ирвина в зависимости от уровня значимости α и значения t.

Если какое-то из полученных расчетных значений λ_t оказывается больше табличного $\lambda_{кр}$, то соответствующий уровень ряда y_t считается аномальным. Чаще всего его заменяют, как было отмечено ранее, средним арифметическим двух соседних с ним уровней.

После исследования аномальных уровней ряда можно перейти к первой задаче второго этапа, а именно провести анализ структуры уровней ряда, используя коэффициенты автокорреляции и ряд других методов по выявлению трендовой составляющей.

4.3. Исследование структуры уровней временного ряда

Предположим, что временной ряд содержит тренд и циклическую или сезонную составляющую. Тогда очевидно, что значение каждого уровня ряда зависит от значений предыдущий уровней.

Автокорреляцией уровней ряда называется корреляционная зависимость между последовательными уровнями ряда.

Для определения силы зависимости между уровнями используется аналогичный парному линейному коэффициенту корреляции коэффициент автокорреляции, для нахождения которого используются уровни исходного ряда и уровни этого же ряда, сдвинутые на несколько шагов во времени.

Формула для расчета коэффициента автокорреляции имеет вид:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \overline{y}_1)(y_{t-1} - \overline{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \overline{y}_1)^2 \sum_{t=2}^{n}(y_{t-1} - \overline{y}_2)^2}},$$

где $\overline{y}_1 = \frac{1}{n-1}\sum_{t=2}^{n} y_t, \qquad \overline{y}_2 = \frac{1}{n-1}\sum_{t=2}^{n} y_{t-1}.$

Коэффициент автокорреляции, приведенный выше, иногда называют коэффициентом автокорреляции уровней ряда первого порядка, так как он измеряет зависимость между соседними уровнями ряда y_t и y_{t-1}.

Также можно определить коэффициент автокорреляции второго порядка, который будет характеризовать силу связи между уровнями y_t и y_{t-2} по формуле:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \bar{y}_3)(y_{t-2} - \bar{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \bar{y}_3)^2 \sum_{t=3}^{n}(y_{t-2} - \bar{y}_4)^2}},$$

где $\bar{y}_3 = \frac{1}{n-2}\sum_{t=3}^{n} y_t, \qquad \bar{y}_4 = \frac{1}{n-2}\sum_{t=3}^{n} y_{t-2}.$

Аналогично можно определить коэффициенты автокорреляции и более высоких порядков, сдвигая уровни исходного ряда на три и более шагов (периодов). Число шагов (периодов), на которое сдвигается исходный ряд и затем рассчитывается коэффициент автокорреляции, называется лагом. Очевидно, что с увеличением лага уменьшается число пар значений, по которым рассчитывается коэффициент автокорреляции. Для обеспечения статистической достоверности рекомендуется не использовать лаг, значение которого будет больше $n/4$.

Коэффициент автокорреляции обладает почти всеми свойствами линейного коэффициента корреляции. Он также изменяется от -1 до 1, характеризует только линейную связь между текущими и предыдущими уровнями ряда. Если связь между уровнями имеет нелинейный характер, то коэффициент автокорреляции может приближаться или быть равным нулю.

Однако между коэффициентами есть и существенные отличия. Знак коэффициента автокорреляции не определяет возрастающую или убывающую тенденцию в уровнях ряда. Например, коэффициент автокорреляции может иметь положительное значение, а временной ряд содержать убывающую тенденцию.

Коэффициенты автокорреляции помогают выявить структуру ряда на предмет наличия трендовой и сезонной (циклической) составляющей с помощью так называемой «автокорреляционной функции». Автокорреляционная функция временного ряда — это функция, значениями которой являются коэффициенты автокорреляции уровней первого, второго и т. д. порядков. Таким образом, аргументом этой функции является величина лага. График автокорреляционной функции называется коррелограммой.

По значениям автокорреляционной функции можно сделать следующие выводы. Если значение коэффициента автокорреляции первого порядка существенно превышает значения этих коэффициентов других порядков, то исследуемый временной ряд содержит только трендовую составляющую. Если коэффициент автокорреляции по-

рядка τ больше коэффициентов других порядков, то ряд содержит сезонные или циклические колебания с периодом τ. И, наконец, если ни один из коэффициентов автокорреляции любого порядка существенно не превышает значения коэффициентов других порядков, то либо ряд не содержит тренда, либо содержит сильный нелинейный тренд, который требует дополнительного анализа.

Таким образом, автокорреляционная функция может быть использована для выявления наличия или отсутствия во временном ряде трендовой и сезонной (циклической) составляющих.

Для выявления наличия трендовой составляющей существуют также другие методы.

Критерий Фостера-Стюарта

Наличие или отсутствие тренда во временном ряду можно проверить по методу Фостера-Стюарта, а также по методу медиан.

Метод Фостера-Стюарта основывается на проверке статистической гипотезы о случайности ряда со следующими пунктами общей схемы.

Основная гипотеза H_0 говорит о том, что тренд отсутствует. Критерием является случайная величина, распределенная по закону Стьюдента. Выборочное значение критерия $t_в$ определяется по формуле:

$$t_в = \frac{D}{\sigma_D},$$

где σ_D – среднеквадратическая ошибка величины D, значения которой заданы в таблице стандартных ошибок σ_D.

Величина D определяется по следующему правилу. Вначале каждый уровень ряда, начиная со второго, сравнивается со всеми предыдущими для нахождения значений m_t и l_t, а именно $m_t = 1$, если y_t строго больше всех предшествующих уровней ряда, в противном случае $m_t = 0$. А $l_t = 1$, если y_t строго меньше всех предшествующих уровней ряда, в противном случае, также $l_t = 0$. Величина $d_t = m_t - l_t$ и, окончательно,

$$D = \sum_{t=2}^{n} d_t.$$

Критическая точка $t_{кр}$ находится по таблице распределения Стьюдента для заданного уровня значимости α и числа степеней свободы $v = n - 1$.

Если $|t_в| > |t_{кр}|$, то гипотеза об отсутствии тренда отвергается.

Итак, о наличии тренда можно судить по значениям коэффициентов автокорреляции уровней или по результатам проверки гипотезы с применением метода Фостера-Стюарта. Однако самым распространенным приемом при выявлении основной тенденции развития процесса, а также дальнейшем ее анализе и использовании при прогнозировании является сглаживание временного ряда.

В результате применения различных приемов сглаживания происходит замена фактических уровней ряда расчетными, которые колебаниям подвержены в меньшей степени. Это позволяет более четко проявиться тенденции развития.

Методы сглаживания условно можно разделить на два класса, в основе применения которых используются два разных подхода: аналитический и алгоритмический.

Аналитическое и алгоритмическое выравнивание временного ряда

При аналитическом подходе допускается возможность задания общей функции, которая описывает регулярную, неслучайную составляющую — тренд. Иногда этот метод называют моделированием тенденции временного ряда.

Для построения трендов чаще всего применяются следующие функции:

– линейный тренд: $\hat{y}_t = a + b \cdot t$;

– гипербола: $\hat{y}_t = a + \frac{b}{t}$;

– экспоненциальный тренд: $\hat{y}_t = \mathrm{e}^{a+b \cdot t}$ (или $\hat{y}_t = a \cdot b^t$);

– степенная функция: $\hat{y}_t = a \cdot t^b$;

– полиномы различных степеней: $\hat{y}_t = a + b_1 \cdot t + b_2 \cdot t^2 + \ldots + b_m \cdot t^m$.

В каждой из перечисленных функций в качестве независимой переменной используется время, а в качестве зависимой — значения фактических уровней. Поэтому функцию (тренд) можно рассматривать в качестве аналога уравнения регрессии. Тогда неизвестные коэффициенты могут быть определены с использованием метода

наименьших квадратов. Для некоторых нелинейных трендов предварительно проводят стандартную процедуру их линеаризации.

После применения метода наименьших квадратов и нахождения статистических оценок неизвестных коэффициентов определяются сглаженные уровни временного ряда. Для этого требуется лишь подставить соответствующее значение временного показателя в полученное уравнение, представленное в явном аналитическом виде.

При применении алгоритмического подхода не предполагается описание с помощью единой функции динамики неслучайной составляющей (тренда). Основными методами, использующими данный подход, являются методы сглаживания временных рядов с помощью скользящей средней. Применение скользящей средней позволяет не только выявить тенденцию в исследуемом процессе, но и сгладить не только случайные, но и периодические колебания.

Кроме определения существования, а также выделения тренда и сезонной компоненты на втором этапе применяют методы исключения тенденции. Исключить трендовую составляющую из временного ряда необходимо при решении задач пятого этапа, т. е. задач исследования взаимосвязи двух и более временных рядов. Если это не сделать, то за счет существования тренда в обоих рядах взаимосвязь между ними может существенно усилиться, что не будет соответствовать действительности.

Ранее был кратко рассмотрен метод моделирования тенденции временного ряда (аналитическое выравнивание). Для получения модели, адекватной действительности, которая в дальнейшем может быть использована для прогнозирования, необходимо выполнение определенных условий. В частности, не должна существовать корреляция между значениями случайной составляющей, так называемая «корреляция в остатках». Анализ случайной составляющей — это задача третьего этапа.

4.4. Автокорреляция в остатках, критерий Дарбина-Уотсона

Автокорреляция уровней ряда, описанная ранее, не ведет ни к каким ложным выводам, а рассматривается как обычное явление, характерное для временных рядов. Иначе обстоит дело с появлением корреляции значений случайной составляющей. Ее наличие может привести к ложным, не соответствующим действительности результатам при исследовании и тем более при прогнозировании. Это свя-

зано с тем, что в случайной составляющей уровней ряда не должно быть никаких закономерностей, в противном случае невозможно применение при исследовании ряда методов, в том числе и метода наименьших квадратов, одной из предпосылок которого является независимость остатков.

Причины возникновения автокорреляции могут быть различны, разной природы. Например, автокорреляция в остатках может появиться из-за ошибок в исходных данных при сборе информации, ошибок измерения признаков. Кроме этого, автокорреляция может возникнуть и в результате неправильной спецификации модели, а именно — в модель может быть не включен признак (фактор), который является значимым для данного процесса. Тогда его влияние будет отражено в случайной составляющей, т. е. в остатках, которые при этом условии будут автокоррелированы.

Отдельно может быть рассмотрена автокорреляция, которая возникла в результате неправильной спецификации функциональной формы модели. В этом случае прежде всего необходимо изменить форму модели, что, возможно, сразу уберет и автокорреляцию в остатках.

Один из самых распространенных методов определения наличия автокорреляции в остатках — это проверка статистической гипотезы с применением критерия Дарбина-Уотсона.

При проверке этой гипотезы в качестве выборочного значения критерия принимается величина, определяемая как отношение суммы квадратов разностей последовательных значений случайной составляющей к остаточной сумме квадратов по модели трендовой составляющей временного ряда:

$$d = \frac{\sum_{t=2}^{n} \left(\varepsilon_t - \varepsilon_{t-1} \right)^2}{\sum_{t=1}^{n} \varepsilon_t^2}.$$

Нетрудно показать, что при достаточно больших значениях n между значением критерием Дарбина-Уотсона d и коэффициентом автокорреляции остатков первого порядка r_1 существует следующее равенство:

$$d = 2 \cdot \left(1 - r_1 \right).$$

Отсюда ясно, что, если $r_1 = 1$, т. е. существует полная положительная корреляция в остатках, то значение критерия $d = 0$. При $r_1 = -1$,

т. е. при наличии полной отрицательной автокорреляции в остатках, значение критерия $d = 4$. При отсутствии автокорреляции в остатках, т. е. когда $r_1 = 0$, значение критерия $d = 2$. Таким образом $0 \le d \le 4$.

После нахождения выборочного значения для исследуемого временного ряда по специальным таблицам для заданного числа наблюдений n, числа независимых факторов трендовой модели m и уровня значимости α определяют две критические точки критерия Дарбина-Уотсона: d_L и d_U. Найденные точки делят весь отрезок от 0 до 4 на пять частей. Решение о наличии или отсутствии автокорреляции в остатках принимается в зависимости от того, в какую часть отрезка попадет значение критерия по следующему правилу:

$0 < d < d_L$ – есть положительная автокорреляция остатков;

$d_L < d < d_U$ – зона неопределенности;

$d_U < d < 4 - d_U$ – автокорреляция остатков отсутствует;

$4 - d_U < d < 4 - d_L$ – зона неопределенности;

$4 - d_L < d < 4$ – есть отрицательная автокорреляция остатков.

Попадание в зону неопределенности означает, что критерий Дарбина-Уотсона не может дать четкого ответа о наличии или отсутствии автокорреляции. Однако на практике в этом случае считают, что автокорреляция существует.

Критерий Дарбина-Уотсона имеет несколько ограничений по применению, а именно, он неприменим к рядам, в которых независимые переменные трендовой модели являются лаговыми переменными результативного признака. Этот критерий выявляет автокорреляцию только первого порядка, и наконец, наиболее достоверные результаты получаются только для больших выборок.

После проведения трех первых этапов анализа временного ряда переходят к решению задач одного из самых важных этапов анализа — четвертого, где рассматриваются возможные методы прогнозирования изучаемого процесса.

ЗАКЛЮЧЕНИЕ

В данной работе была рассмотрена та часть математико-статистического исследования, которая связана, в основном, с применением методов многомерного математико-статистического анализа.

Эти методы непрерывно развиваются и совершенствуются, расширяется их область применения. Особенно успешно происходит применение этих методов с использованием новейших средств вычислительной техники и программных средств (пакетов прикладных программ — SPSS, Statistika и т. д.)

Однако не стоит забывать, что выводы, прогнозы и рекомендации, полученные после математико-статистического исследования, будут точны и надежны только тогда, когда, во-первых, выбор соответствующего метода анализа научно обоснован и, во-вторых, этот метод правильно применен с учетом всех ограничений и допущений.

Список литературы

Основная:

1. Большакова, Л.В., Примакин, А.И., Яковлева, Н.А. Математико-статистические методы обработки экспериментальных данных при проведении научных исследований: методические рекомендации: в 3-х частях. Часть 1. – Санкт-Петербург: Изд-во СПб ун-та МВД России, 2014. – 92 с.

2. Большакова, Л.В., Яковлева, Н.А. Математико-статистические методы обработки экспериментальных данных при проведении научных исследований: методические рекомендации: в 3-х частях. Часть 2. – Санкт-Петербург: Изд-во СПб ун-та МВД России, 2018. – 68 с.

3. Литвиненко, А.Н., Большакова, Л.В. Методика применения кластерного анализа при выполнении выпускных квалификационных работ слушателями Санкт-петербургского университета МВД России // Вестник Санкт-Петербургского университета МВД России. – 2020. – № 1 (85). – С. 208–218.

Дополнительная:

1. Большакова, Л.В., Примакин, А.И., Яковлева, Н.А. Применение кластерного и дискриминантного анализов в процессе обработки и интерпретации статистических данных при обеспечении экономической и информационной безопасности хозяйствующего субъекта // Вестник Санкт-Петербургского университета МВД России. – 2014. – № 2 (62). – С. 148–156.

2. Винюков, И.А. Многомерные статистические методы: учебное пособие. – Москва: Финансовый университет, 2014. – 90 с.

3. Кремер, Н.Ш. Теория вероятностей и математическая статистика. – Москва: ЮНИТИ-ДАНА, 2012. – 550 с.

4. Симчера, В.М. Методы многомерного анализа статистических данных: учебное пособие. – Москва: Финансы и статистика, 2008. – 395 с.

5. Васильева, Э.К., Юзбашев, М.М. Выборочный метод в социально-экономической статистике: учебное пособие. – Москва: Финансы и статистика, ИНФРА-М, 2010. – 254 с.

6. Чубукова И.А. Курс лекций «Data Mining» // Интернет-университет информационных технологий. URL: www.intuit.ru/department/database/datamining/ (дата обращения: 18.06.2023).

Приложение 1. Критические точки распределения Стьюдента

Число степеней свободы υ	Уровень значимости α (двусторонняя критическая область)					
	0,1	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	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	3,75	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	5,04
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	3,37
Число степеней свободы υ	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Приложение 2. Критические точки распределения Пирсона χ^2

Число степеней свободы υ	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Приложение 3. Критические точки метода Ирвина

Число измерений *n*	Уровень значимости	
	q=0,05	*q*=0,01
2	2,8	3,7
3	2,2	2,9
10	1,5	2,0
20	1,3	1,8
30	1,2	1,7
50	1,1	1,6
100	1,0	1,5
400	0,9	1,3
1000	0,8	1,2

Приложение 4. Значения статистик Дарбина-Уотсона $d_L d_U$ при 5%-м уровне значимости

n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0,61	1,40								
7	0,70	1,36	0,47	1,90						
8	0,76	1,33	0,56	1,78	0,37	2,29				
9	0,82	1,32	0,63	1,70	0,46	2,13				
10	0,88	1,32	0,70	1,64	0,53	2,02				
11	0,93	1,32	0,66	1,60	0,60	1,93				
12	0,97	1,33	0,81	1,58	0,66	1,86				
13	1,01	1,34	0,86	1,56	0,72	1,82				
14	1,05	1,35	0,91	1,55	0,77	1,78				
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,85	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,99
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83

Приложение 5. Критические точки распределения Фишера-Снедекора при 5%-м уровне значимости

k1 \ k2	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48

Учебное издание

Большакова Людмила Валентиновна,
кандидат физико-математических наук, доцент;
Сибаров Константин Дмитриевич,
кандидат технических наук, доцент;
Яковлева Наталья Александровна,
кандидат психологических наук

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ ПРИ ПРОВЕДЕНИИ НАУЧНЫХ ИССЛЕДОВАНИЙ

Методические рекомендации

В 3-х частях

Часть 3

Редактор *Свикша Н.О.*
Компьютерная верстка *Свикша Н.О.*
Дизайн обложки *Шеряй А.Н.*

ISBN 978-5-91837-752-9

EDN: DAHTQH

Подписано в печать 07.08.2023. Формат 60×84 $^{1}/_{16}$
Печать цифровая 5,75 п. л. Тираж 50 экз. Заказ № 70/23

Отпечатано в Санкт-Петербургском университете МВД России
198206, Санкт-Петербург, ул. Летчика Пилютова, д. 1